MULTIOMIC APPROACHES FOR ASSESSING THE ROLE OF NATURAL MICROBIAL COMMUNITIES IN NITROUS OXIDE EMISSION FROM MIDWESTERN AGRICULTURAL SOILS

A Dissertation
Presented to
The Academic Faculty

by

Luis Humberto Orellana Retamal

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY in the
SCHOOL OF CIVIL AND ENVIRONMENTAL ENGINEERING

Georgia Institute of Technology
August 2017

COPYRIGHT © 2017 BY LUIS ORELLANA

MULTIOMIC APPROACHES FOR ASSESSING THE ROLE OF NATURAL MICROBIAL COMMUNITIES IN NITROUS OXIDE EMISSION FROM MIDWESTERN AGRICULTURAL SOILS

Approved by:

Dr. Konstantinos T. Konstantinidis,
Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Joe Brown
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Spyros G. Pavlostathis
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Joel E. Kostka
School of Biology and Earth & Atmospheric Sciences
Georgia Institute of Technology

Dr. Jim C. Spain
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Frank E. Löffler
School of Civil and Environmental Engineering
University of Tennessee

Date Approved: April 28, 2017

To my Grandmother Noelia and Mother Cecilia

ACKNOWLEDGEMENTS

I would like to thank Dr. Kostas Konstantinidis for the opportunity to be part of his research group and for all the support, patience and guidance throughout the years that made this work possible. I am deeply grateful for the research and collaboration opportunities he provided me, which have helped me develop my academic career. I would like to thank my committee members Dr. Jim Spain, Dr. Spyros Pavlostathis, Dr. Joe Brown, Dr. Joel Kostka, and Dr. Frank Loëffler for their always insightful feedback and support. Also thanks to Dr. Rob Sanford and Dr. Joanne Chee-Sanford for the incredible research experience in Illinois. I would also like to thank Fulbright and Conicyt for providing me the financial support during my first 3 years at Gatech. I would also like to thank all the current members of Kostas Lab, but a special thanks to Alejandro, Natasha, Despina, and Miguel who were of immense help when I started my studies at Gatech (Seriously, you have no idea!). To all the friends I have made in Atlanta along the way – thanks for making this experience more enjoyable. I would like to thank my friends in Chile, Sofía, María José, and Cecilia that, even though are far away, are always supportive. I would like to thank the long list of my family members, but especially my mother Cecilia, sister Camila, and father Humberto for their incredible and unconditional support. I would also like to thank my family in the U.S., Julie, Kent and Lyndon, for making feel at home every time I visit. Finally, thank you to my wonderful fiancée Alissa, who somehow agreed to be part of this journey even before coming to Atlanta and never ceases to support and help me when I need it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Anthropogenic activities such as fossil fuel consumption and industrial nitrogen (N) fixation processes have increased the N inputs into the environment. Even though the central role of microbes in the cycling of N is recognized, the identification and diversity of these microbial pathways in agricultural soils are still lacking. This scarcity of information limits the development of more accurate, predictive models of N-flux including the role of microbes in the generation and consumption of important nitrogenous greenhouse gases (e.g., nitrous oxide, N_2O). The advent of new high-throughput nucleic acid sequencing technologies allows nowadays the exploration of soil microbial communities that were previously insufficiently studied based on cultivation and PCR approaches. In this work, we integrated experimental data and bioinformatics approaches to identify and quantify indigenous soil microorganisms participating in the cycling of nitrogen in agricultural fields, particularly those involved in the generation and consumption of N_2O. We developed a new bioinformatic approach, called ROCker, to accurately detect target genes and transcripts in complex short-read metagenomes and metatranscriptomes, which offered up to 60-fold lower false discovery rate compared to the common strategy of using e-value thresholds. Using ROCker, we found an unexpectedly high abundance of nitrous oxide reductase genes, the only known biological sink of N_2O, in soil and aquatic environments. In two particular soil types that typify the Midwest cornbelt, we show that microbial communities are remarkably stable across the year compared to other environments except during nitrogen fertilization events, which

stimulate the activity of novel nitrogen-utilizing *Nitrospirae* and *Thaumarchaeota* taxa. Lastly, we assessed the predictive power of omic techniques in estimating nitrification process rates in incubated soil mesocosms and found high correlations between target transcripts and experimentally measured nitrification activity, providing new molecular means to measure microbial activity *in-situ*. These findings have implications for understanding the diversity and dynamics of natural microbial communities controlling the N cycle in soils.

CHAPTER 1. INTRODUCTION

1.1 BACKGROUND

1.1.1 Nitrogen Cycle and Nitrogenous Greenhouse Gases

About 78% of earth's atmosphere is composed of dinitrogen (nitrogen gas, N_2), a chemical form of nitrogen (N) inaccessible to most organisms. Even though a strong triple bond maintains the two atoms of nitrogen together in the N_2 molecule, some microorganisms developed an enzymatic strategy to convert or 'fix' N_2 into a form that can be readily used by most life forms. This fixed nitrogen (mostly in the form of ammonium, NH_4^+) can be transformed into organic compounds such as amino acids and nucleotides, in a process that is essential for generating the building blocks of a cell. On the other hand, the reduced forms of N (e.g., NH_4^+) can be oxidized to nitrite (NO_2^-) and nitrate (NO_3^-) and used as an energy source for some organisms, in a process called nitrification. The oxidized forms of nitrogen (e.g., NO_3^-) can be sequentially used as electron acceptors by many organisms, generating in turn, reduced forms of N, in a process called denitrification. The tight coupling of all these processes is known as the N-cycle and is mostly catalyzed by microorganisms such as bacteria, archaea and fungi.

Anthropogenic activities are severely impacting the nitrogen (N) cycle. The industrial fixation of N, by means of the Haber-Bosch process (i.e., $N_2 + 3\ H_2 \rightarrow 2\ NH_3$) introduces more than 100 Tg of reactive nitrogen

per year, mostly as ammonium-based (NH_4^+) fertilizers (1). In addition, the cultivation of crops and deposition of NO_x from fossil fuel combustion adds about 40 and 25 Tg, respectively, of extra N into the environment every year (2). These anthropogenic activities have led to increased emissions of nitrous oxide gas (N_2O). Despite the fact that only a small fraction of the applied N is emitted as N_2O, close to 7 Tg, or 40%, of total N_2O emissions arise from agricultural activities (3).

Nitrous oxide is a potent greenhouse gas and catalyst of ozone depletion in the stratosphere (2, 4, 5) and it is mainly generated as a result of biotic pathways (6). For instance, diverse microbial populations performing complete denitrification (NO_3^-/NO_2^- N_2), nitrifier denitrification (NH_4^+ N_2), nitrification (NH_4^+ NO_2^-/ NO_3^-), incomplete denitrification (NO_3^- $/NO_2^-$ N_2O/N_2) and ammonification (e.g., dissimilatory nitrate reduction to ammonium, NO_3^- NH_4^+), can contribute to the emission of N_2O as a result of these reactions. Furthermore, other coupled microbial-abiotic sources of N_2O have been recognized, such as the chemical reduction of nitrite in the presence of hydroxylamine ($NH_2OH + NO_2^- + 2H+$ $N_2O + 2H_2O$). Also, it has been proposed that ferrous iron generated by ferric iron-reducing bacteria can react with high concentrations of nitrite to produce N_2O abiotically, in a process termed chemo-denitrification (7).

1.1.2 New Contributors to Nitrification

The addition of N to soils is a conventional agricultural practice for obtaining higher crop yields. Synthetic fertilizers (e.g., a mixture of urea

and ammonium-nitrate) typically attach to clay and organic materials, facilitating their retention in soils, and hence, uptake by plants. However under aerobic conditions, ammonia-oxidizing microorganisms can rapidly convert ammonium to nitrite and nitrate, generating nitrous oxide gas (dinitrogen oxide, N_2O) as a by-product (8). In addition, the negative charge of nitrate repels it from the soil matrix leaching it out of the root zone and therefore producing massive N losses for agriculture and contamination of drinking water and eutrophication of water bodies by nitrates.

During nitrification, the ammonium oxidation to nitrite is considered the first and rate-limiting step (9) and it was originally described by Sergei Winogradsky (Equation 1). The sequential oxidation of nitrite to nitrate has been historically a much less studied process, despite its key role in generating oxidized N forms for denitrification (Equation 2).

$$\mathrm{NH_4^+ + 1.5O_2 \rightarrow NO_2^- + H_2O + 2H^+} \left(\Delta \mathrm{G^{0'}} = -274.7\ \mathrm{kJ\ mol^{-1}}\right) (1)$$

$$\mathrm{NO_2^- + 0.5O_2 \rightarrow NO_3^-} \qquad \left(\Delta \mathrm{G^{0'}} = -\ 74.1\ \mathrm{kJ\ mol^{-1}}\right) (2)$$

Even though a reduced group of autotrophic soil *Betaproteobacteria* were originally described as the drivers for this two-step sequential oxidation process, the exploration of the microbial diversity in soils and oceans has revealed a much more complicated picture. For instance, the *Gammaproteobacterium Nitrosococcus oceani* first expanded the taxonomic repertoire of bacteria involved in nitrification (10). In addition, an early metagenomic survey of the Sargasso Sea (11), along with the cultivation of an ammonia oxidizing archaeon

(12) from an aquarium tank, first detected ammonium oxidation in the archaeal domain of life. Subsequently, it was found that in many environments, the abundances of the ammonia monooxygenase (*amoA*) genes belonging to ammonia-oxidizing archaea (AOA) often exceed their bacterial counterparts (AOB), suggesting a major role of archaea in nitrogen cycle in soils and oceans (13, 14) and importance for N_2O emissions (15).

Further, even though nitrification has been historically studied as a two-step process, thermodynamics calculations have proposed that this process could occur as a single reaction:

$$\mathrm{NH_4^+ + 2O_2 \rightarrow \ NO_3^- + H_2O + 2H^+ \left(\Delta G^{0\prime} = -348.9\ kJ\ mol^{-1}\right)} \quad (3)$$

In fact, previous studies have proposed that a single organism living under limited substrate influx and forming microbial aggregates or living in biofilms would be able to perform complete ammonia oxidation (“Comammox”) (16). This hypothetical microbe would take advantage of the larger free energy obtained from the single step process (equation 3) and hence, compete with those microbes performing the two-step process (equation 1 and 2). Such a microorganism remained elusive until the end of 2015. Two parallel ground-breaking publications lead by M. Jetten (17) and M. Wagner (18) reported independent enrichments containing bacteria affiliated with the *Nitrospira* genus, which were able to oxidize ammonia to nitrate. By performing metagenomics and contig binning, they were able to reconstruct the genomes for the *Candidatus* Nitrospira nitrosa, *Candidatus* Nitrospira nitrificans (17), and *Candidatus* Nitrospira inopinata (18). Surprising for that time, these chemolitoautotrophic

organisms encoded both pathways for ammonia and nitrite oxidation in their genomes, but the sequences of the underlying *amoA* and the hydroxylamine dehydrogenase (*hao)* proteins responsible for ammonia oxidation were phylogenetically divergent from previously described proteins. This novel reaction was observed in enrichment cultures under laboratory conditions. Thus, further work is necessary to test the importance and relevance of Comammox *in situ*.

1.1.3 Unaccounted Nitrous Oxide (N_2O) Gas Sinks

As far as the denitrification part of the nitrogen cycle is concerned, the recent discovery of an unaccounted group of microorganisms encoding an “atypical” nitrous oxide reductase (*nosZ)* is advancing our understanding of biological N_2O sinks in the environment. NosZ, commonly found in the genomes of complete denitrifiers, is the key enzyme and only known biological sink for N_2O. The phylogenetic analysis of NosZ sequences retrieved from sequenced genomes showed the existence of two distinct clades revealing an evolutionary distinct origin for these enzymes. Interestingly, PCR primers commonly used for studying the diversity and activity (i.e., expression) of *nosZ* only amplified sequences from one clade of the tree or the “typical” sequences commonly found in complete denitrifier organisms (19, 20). Remarkably, Sanford (19) and collaborators found that microorganisms encoding “atypical” *nosZ* can reduce N_2O to N_2. Therefore, it was proposed that the analysis of both typical and atypical *nosZ* genes is necessary to understand and predict the emission of N_2O in the environment.

1.1.4 Metagenomics as an Approach for Studying Natural Microbial Communities

The advent of sequencing technologies in recent years has allowed the exploration and identification of microbial communities previously elusive to culturing techniques. These communities were first accessed by sequencing or randomly cloning DNA fragments in bacterial vectors that were later screened for target metabolic functions (21). Even though the early sequencing of random environmental DNA (i.e., metagenomics) revealed unexpected microbial diversity in marine environments (22) or soils (23), it was limited to kilobases of information from clones per study, in addition to laborious work and high economic cost. However, the quest for economically sequencing the entire human genome introduced a new approach. Instead of cloning and sequencing individual clones, a semiautomatic approach of sequencing of random fragments (i.e., “shotgun”) of DNA was invented. The latter technology, with additional optimizations and technological breakthroughs, further expanded the frontiers of DNA sequencing, providing nowadays terabases in short DNA fragments (50-500bp) at a much lower labor and cost.

1.1.5 Current Challenges for Studying Genes and Population Genomes in Short-Read Metagenomes

Metagenomic or metatranscriptomic approaches commonly used to study microbial communities in clinical or environmental samples are

currently challenged by the lack of bioinformatic approaches and several technical challenges. For instance, similarity search algorithms (e.g., BLAST) allow the identification of protein or nucleotide homologs by evaluating the alignment of a query sequence to a known reference. The summary of the alignment score is then used as the likelihood of finding a match in a specific database by chance (e-value), and a score below this threshold e-value is commonly used as a significant match (i.e., query sequence is homologous to its matching reference sequence). Although the use of e-values represents an efficient method for selecting matches, these values are often arbitrarily chosen and do not necessarily guarantee true homology. As a consequence, the rate of false positive (i.e., incorrectly identified, FP) or false negative (i.e., incorrectly rejected, FN) matches have been rarely assessed nor rigorously evaluated. Thus, in metagenomes of highly diverse communities such as those obtained from soils, ocean or the human gut, the use of arbitrary cutoffs can result in an undetermined number of FPs. These difficulties are increased when genes or proteins of interest share highly conserved domains or motifs such as metal binding or ATP-hydrolyzing domains, which can potentially retrieve a high fraction of FP matches. Furthermore, these tools were not originally designed to be used with short-read sequences of the current sequencing technologies, hence, the robustness of this approach for short-read metagenomics remains unclear.

In **chapter 2**, we introduce a bioinformatic approach called ROCker that accurately identifies metagenomic or metatranscriptomic reads encoding a target protein family of interest by determining the most-discriminating bitscore thresholds across the sequence of the protein. Briefly, simulated metagenome-like datasets of known compositions are generated from sequenced microbial genomes encoding the target protein(s). These datasets are then used as a training set for generating a profile of most discriminating, position-specific, bitscores (a value reflecting the score of the alignment) across the target protein alignment. The calculated bitscore values maximize the recovery of true positive and minimizes false positive matches based on the receiver operating characteristic (ROC) curve.

We showcased ROCker using two highly similar proteins: the ammonia monooxygenase subunit a gene (*amoA*) and the particulate methane monooxygenase (*pmoA*), which are not typically encoded on the same genomes and are often challenging to distinguish from each other (e.g. share >60% amino acid identity). ROCker showed a 60-fold improvement in false discovery rate (FDR) compared to the use of a fixed e-value threshold (e.g., 10^{-5}) for these proteins. We further evaluated ROCker for studying key nitrogen genes such as nitrous oxide reductase (*nosZ*) and nitrite reductase (*nirK*) in available metagenomes representing soil and marine sediment ecosystems. ROCker quantified the abundance of N-genes much more precisely than the common practice of fixed e-

values by excluding matching reads encoding shared/overlapping functional domains. The analysis of soil metagenomes showed a higher ratio of *nirK*/*nosZ* for terrestrial samples relative to marine sediments. These findings are consistent with the hypothesis that in some environments, a high fraction of organisms capable of performing denitrification do not possess the genetic potential to reduce N_2O to N_2 gas. We showed that the use of ROCker in clinical and environmental surveys expands the molecular toolbox for study microbial communities and their activities *in situ* by providing an efficient and accurate pipeline for quantifying target genes in short-read metagenomes.

1.1.6 Examining the Diversity of N Cycle Genes and Dynamics of Indigenous Microbial Communities in Agricultural Soils

Even though agricultural soil ecosystems are characterized by a dynamic interplay between complex biotic and abiotic processes that maintain the cycling of nutrients in the ecosystem (21), little is known about the impact of seasonal agricultural activities on natural microbial populations. This paucity in information limits our understanding and predictive modeling capability of the microbial pathways involved in the cycling of key nutrients such as carbon and nitrogen. Moreover, the recent discovery of the Comammox *Nitrospira* encoding all necessary enzymes to perform complete nitrification have initiated new interest in the N cycle but the prevalence and activity of these organisms in soil ecosystems remains essentially unknown. The ecological interactions and relative

activities of AOA, AOB, NOB and Comammox *Nitrospira* in soils represent currently a hot topic of research due to their obvious importance in nitrogen cycling and greenhouse gas emissions (e.g., N_2O production). Thus, the identification and tracking of natural communities responding to N fertilization can provide new insights and models of the generation and consumption of N_2O in agricultural soils during the growing season.

In agricultural soils receiving large inputs of nitrogen, determining the abundance and diversity of *nosZ* genes is important for accurate prediction of microbial activity potentially reducing N_2O to N_2. To circumvent the limitations for detecting *nosZ* genes in the environment using PCR, in **chapter 3** we analyzed short-read metagenomes from agricultural soils and different locations. Using the ROCker methodology described above, we queried soil metagenomes obtained from two agricultural soils with different physicochemical properties and a long history of fertilization practices located in the U.S. Midwest. Our analyses found that 70%, or more, of the total *nosZ* reads detected were classified as atypical, underscoring that previous gene surveys have underestimated *nosZ* abundance in soils. These results showed that atypical *nosZ* genes were likely missed in previous PCR-based surveys due to the low nucleotide identity between the two *nosZ* sequences (60.9%+/- 8.2. A high abundance of atypical *nosZ* reads was also detected in soil metagenomes obtained from distant geographic locations and representing a variety of biomes. In conclusion, the high abundance of previously unaccounted

atypical *nosZ* genes in soils suggests that denitrifiers that harbor atypical *nosZ* genes may contribute more than previously assumed to the reduction of N_2O to innocuous N_2 gas.

Using a metagenomic approach, in **chapter 4** we described the diversity and temporal dynamics of natural microbial communities present in two agricultural soils of contrasting soil textures representative of the US Midwest. Samples were collected at two different depths during the growing season in 2012 and showed a remarkably stable genetic potential. In fact, differences in the gene content were evident between soil layers and sites rather than the sampling points. For instance, genes related to light stress, DNA repair and nutrient uptake were abundant in the top layer whereas nitrogen and divergent archaeal metabolism genes were characteristic of the deeper soil samples (>2-fold, p-adj<0.05). In addition, only ~20% of the recovered metagenomic bin populations showed more than 2-fold change in abundance between sampling points. Among the latter bins, deep branching *Thaumarchaeota* and Comammox *Nitrospira* showed up to 5-fold abundance increase upon the addition of nitrogen fertilizer in the sandy site. The bins encoded all the genes necessary to perform complete nitrification and shared over 66% AAI with the recently isolated Comammox bacterium *Nitrospira inopinata* (18). The *Thaumarchaeota* bins were affiliated to the previously described I.1b and I.1a clades and shared between ~70% AAI among them. Interestingly, the sharp increase in abundance observed for *Thaumarchaeota* organisms might suggest they respond faster to N-fertilization than expected based on the assumptions that they are K-strategists (slow

growers). Similar results to those reported above for individual populations were observed N-cycle genes encoded by metagenomics reads. Altogether, these results show stable microbial communities and novel key organisms responding to agricultural management and propose a much broader impact for ammonia oxidizing *Thaumarchaeota* and *Nitrospira*.

1.1.7 Omic Approaches for Assessing Microbial Activity in situ

Even though our results showed that metagenomes (DNA level) represent reliable means to study microbial communities in soils, RNA and proteins may reflect different levels of information closer to microbial activity. While gene sequences offer a comprehensive overview of the genetic potential of microbial communities, it is the mRNA molecules, and especially the proteins, that are closer to activity and enzymatic functions. Thus, short-term microbial responses to external changes (e.g., nitrogen additions) can best be tracked by analyzing the information from DNA, RNA and proteins. Although several recent studies have used metagenomic, metatranscriptomic and metaproteomic approaches on soil samples (22, 23), it is not clear to what degree these different levels of information correlate with each other or with *in situ* measured rates. Most studies to date have used single bacterial or archaeal isolates under laboratory conditions, therefore reflecting a small fraction of the extant microbial diversity and its activity under artificial conditions compared to field conditions. In **chapter 5**, we examined the integration of microbial activity obtained from DNA, RNA and protein abundances in soil microcosms incubated with ammonia. The amendment of nitrogen to incubated Havana soils showed a high nitrification rate

after 2 days of incubation (~2μ NO_3^--N g^{-1} d^{-1}), and ^{15}N-N_2O production from ^{15}N-NH_4^+ oxidation was detected after 1 day of incubation. The tracking of nitrification genes by metatranscriptomics revealed much stronger responses of nitrifier organisms to the treatment compared to metagenomic data. Also, a conspicuous dynamic response from bacterial *amoA* transcripts was observed after 8 days of incubation, whereas their archaeal counterparts showed higher but stable abundance during the incubation time. Similar increased abundance throughout the incubations was observed for hydroxylamine oxidase (*hao)* and nitrite oxidoreductase (*nxrA)* genes mostly phylogenetically affiliated to *Betaproteobacteria* and *Nitrospira*, respectively. Interestingly, *Betaproteobacteria* nitrification genes showed increased abundance whereas Comammox *Nitrospira* transcripts were stable throughout the incubations. Even though fewer peptides were detected compared to genes and transcripts, results derived from metaproteomics showed increased abundance for ATP synthases and transcription proteins in N-amended soils. As expected from measured nitrite oxidation rates and transcripts, NxrB peptides showed increased abundance after 8 days of incubation compared to the control soils. Finally, linear regressions between N_2O generation and bacterial *amoA* transcript abundance showed significant and positive correlations (Pearson R^2>0.95) whereas archaeal *amoA* transcripts correlated less strongly with measured N_2O production or ammonia consumption rates. Thus, the comparative analyses of three omic techniques from the soil incubations showcased the advantages of each

technique, and outlined an approach to use molecular data as a proxy for examining biological processes in soils.

In the following chapters, novel bioinformatic approaches and their application to study the role of microorganisms in the N cycle in soils are presented. Specifically, the discovery and description of novel microbial communities potentially participating in the N cycle and consequently on the generation and consumption of the potent greenhouse gas, N_2O, in agricultural soils is presented in Chapters 2,3, and 4. In addition, the results show that information gathered from the DNA, RNA and proteomes recovered from natural microbial populations can be effectively used to explore key microbial pathways and process rates *in situ* (Chapter 5). The approaches presented here provide new means and recommendations for effectively examining microorganisms and their activities in soils and elsewhere. For instance, the novel microbial diversity presented here can be incorporated in earth models by environmental engineers and scientists interested in predicting microbial sources of consumption of N_2O in natural or engineered systems. Altogether, the results provide new insights into the role of natural communities in N cycling, and provide a more comprehensive picture of the microbial communities involved in N_2O generating and consuming pathways.

1.2 REFERENCES

1. **Fields S**. 2004. Global nitrogen: cycling out of control. Environ Health Perspect **112**:A556–63.

2. **Gruber N**, **Galloway JN**. 2008. An Earth-system perspective of the global nitrogen cycle. Nature **451**:293–296.

3. **Montzka SA**, **Dlugokencky EJ**, **Butler JH**. 2011. Non-CO2 greenhouse gases and climate change. Nature **476**:43–50.

4. **Ravishankara AR**, **Daniel JS**, **Portmann RW**. 2009. Nitrous oxide (N2O): the dominant ozone-depleting substance emitted in the 21st century. Science **326**:123–125.

5. **Fowler D**, **Coyle M**, **Skiba U**, **Sutton MA**, **Cape JN**, **Reis S**, **Sheppard LJ**, **Jenkins A**, **Grizzetti B**, **Galloway JN**, **Vitousek P**, **Leach A**, **Bouwman AF**, **Butterbach-Bahl K**, **Dentener F**, **Stevenson D**, **Amann M**, **Voss M**. 2013. The global nitrogen cycle in the twenty-first century. Philos Trans R Soc Lond, B, Biol Sci **368**:20130164–20130164.

6. **Singh BK**, **Bardgett RD**, **Smith P**, **Reay DS**. 2010. Microorganisms and climate change: terrestrial feedbacks and mitigation options. Nature Reviews Microbiology **8**:779–790.

7. **Kampschreur MJ**, **Kleerebezem R**, **de Vet WWJM**, **van Loosdrecht MCM**. 2011. Reduced iron induced nitric oxide and nitrous oxide emission. Water Research **45**:5945–5952.

8. **Bremner JM**. 1997. Sources of nitrous oxide in soils. Nutrient Cycling in Agroecosystems **49**:7–16.

9. **Kirchman DL**. 2012. Processes in Microbial Ecology. Oxford University Press.

10. **Klotz MG**, **Arp DJ**, **Chain PSG**, **El-Sheikh AF**, **Hauser LJ**, **Hommes NG**, **Larimer FW**, **Malfatti SA**, **Norton JM**, **Poret-Peterson AT**, **Vergez LM**, **Ward BB**. 2006. Complete genome sequence of the marine, chemolithoautotrophic, ammonia-oxidizing bacterium Nitrosococcus oceani ATCC 19707. Appl Environ Microbiol **72**:6299–6315.

11. **Venter JC**, **Remington K**, **Heidelberg JF**, **Halpern AL**, **Rusch D**, **Eisen JA**, **Wu D**, **Paulsen I**, **Nelson KE**, **Nelson W**, **Fouts DE**, **Levy S**, **Knap AH**, **Lomas MW**, **Nealson K**, **White O**, **Peterson J**, **Hoffman J**, **Parsons R**, **Baden-Tillson H**, **Pfannkoch C**, **Rogers Y-H**, **Smith HO**. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304**:66–74.

12. **Könneke M**, **Schubert DM**, **Brown PC**, **Hügler M**, **Standfest S**, **Schwander T**, **Schada von Borzyskowski L**, **Erb TJ**, **Stahl DA**, **Berg IA**. 2014. Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation. **111**:8239–8244.

13. **Stieglmeier M**, **Alves RJE**, **Schleper C**. 2014. The Phylum Thaumarchaeota, pp. 347–362. *In* Rosenberg, E, DeLong, EF, Lory, S, Stackebrandt, E, Thompson, F (eds.), The Prokaryotes. Springer Berlin Heidelberg, Berlin, Heidelberg.

14. **Schleper C**, **Nicol GW**. 2010. Ammonia-oxidising archaea--physiology, ecology and evolution. Adv Microb Physiol **57**:1–41.

15. **Prosser JI**, **Nicol GW**. 2012. Archaeal and bacterial ammonia-oxidisers in soil: the quest for niche specialisation and differentiation. Trends in Microbiology **20**:523–531.

16. **Costa E**, **Pérez J**, **Kreft J-U**. 2006. Why is metabolic labour divided in nitrification? Trends in Microbiology **14**:213–219.

17. **van Kessel MAHJ**, **Speth DR**, **Albertsen M**, **Nielsen PH**, **Op den Camp HJM**, **Kartal B**, **Jetten MSM**, **Lücker S**. 2015. Complete nitrification by a single microorganism. Nature **528**:555–559.

18. **Daims H**, **Lebedeva EV**, **Pjevac P**, **Han P**, **Herbold C**, **Albertsen M**, **Jehmlich N**, **Palatinszky M**, **Vierheilig J**, **Bulaev A**, **Kirkegaard RH**, **Bergen von M**, **Rattei T**, **Bendinger B**, **Nielsen PH**, **Wagner M**. 2015. Complete nitrification by Nitrospira bacteria. Nature **528**:504–509.

19. **Sanford RA**, **Wagner DD**, **Wu Q**, **Chee-Sanford JC**, **Thomas SH**, **Cruz-García C**, **Rodríguez G**, **Massol-Deyá A**, **Krishnani KK**, **Ritalahti KM**, **Nissen S**, **Konstantinidis KT**, **Löffler FE**. 2012. Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils. **109**:19709–19714.

20. **Jones CM**, **Graf DR**, **Bru D**, **Philippot L**, **Hallin S**. 2013. The unaccounted yet abundant nitrous oxide-reducing microbial community: a potential nitrous oxide sink. The ISME Journal **7**:417–426.

21. **Dick RP**. 1992. A review: long-term effects of agricultural systems on soil biochemical and microbial parameters. Agriculture, Ecosystems & Environment **40**:25–36.

22. **Williams MA**, **Taylor EB**, **Mula HP**. 2010. Metaproteomic characterization of a soil microbial community following carbon amendment. Soil Biology and Biochemistry **42**:1148–1156.

23. **Hultman J**, **Waldrop MP**, **Mackelprang R**, **David MM**, **McFarland J**,

Blazewicz SJ, **Harden J**, **Turetsky MR**, **McGuire AD**, **Shah MB**, **VerBerkmoes NC**, **Lee LH**, **Mavrommatis K**, **Jansson JK**. 2015. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature **521**:208–212.

CHAPTER 2. ROCKER: ACCURATE DETECTION AND QUANTIFICATION OF TARGET GENES IN SHORT-READ METAGENOMIC DATASETS

Reproduced with permission from Luis H Orellana, Luis M Rodriguez-R, Kostas T Konstantinidis. *Nucleic Acids Research*. 2016 October 7. Copyright © Oxford University Press on behalf of Nucleic Acids Research.

2.1 ABSTRACT

Functional annotation of metagenomic and metatranscriptomic datasets relies on similarity searches based on e-value thresholds resulting in an unknown number of false positive and negative matches. To overcome these limitations, we introduce ROCker, aimed at identifying position-specific, most-discriminant thresholds in sliding windows along the sequence of a target protein, accounting for non-discriminative domains shared by unrelated proteins. ROCker employs the receiver operating characteristic (ROC) curve to minimize false discovery rate (FDR) and calculate the best thresholds based on how simulated shotgun metagenomic reads of known composition map onto well-curated reference protein sequences and thus, differs from HMM profiles and related methods. We showcase ROCker using ammonia monooxygenase (*amoA*) and nitrous oxide reductase (*nosZ*) genes, mediating oxidation of ammonia and the reduction of the potent greenhouse gas, N_2O, to inert N_2, respectively. ROCker typically showed 60-fold lower FDR when compared to the common practice of using fixed e-values. Previously uncounted "atypical" *nosZ* genes were found to be two times more abundant, on average, than their typical counterparts in most soil

metagenomes, and the abundance of bacterial *amoA* was quantified against the highly-related archaeal *amoA* and particulate methane monooxygenase (*pmoA*). Therefore, ROCker can reliably detect and quantify target genes in short-read metagenomes.

2.2 INTRODUCTION

Omics approaches are commonly applied to the study of microbial communities in a variety of clinical and environmental settings, but numerous technical challenges remain for accurately analyzing short gene sequences recovered from metagenomes or metatranscriptomes (1). Most importantly, several standard bioinformatic tasks rely on widely used similarity search algorithms (e.g., BLAST) that, through the comparison of nucleic or protein sequences to reference databases, allow for the identification of homologous genetic features among millions of unrelated sequences. However, in short-read metagenomes or metatranscriptomes representing diverse microbial communities (e.g., those of soils, oceans, or the human gut), the rate of false positive (i.e., incorrectly identified, FP) or false negative (i.e., incorrectly rejected, FN) matches obtained from similarity searches are rarely addressed or quantified. An important underlying cause for FP and FN matches is the use of thresholds for a match based on a fixed e-value, a statistical parameter that reflects the number of expected matches by chance but not necessarily true homology. Although the use of e-values represents an efficient strategy for selecting matches, it can result in a substantial number of false positives, especially for protein sequences that share functional domains or motifs. Only

lately, these limitations have received adequate attention but mostly for taxonomic assignment purposes (2, 3).

Recently, we employed the receiver operating characteristic curve (ROC) approach to refine the results of similarity searches and calculate a reliable, fixed bitscore value across the sequence of the target gene that maximizes the sensitivity (true positive rate) and specificity (true negative rate) for detecting short-gene fragments encoding nitrous oxide reductase (*nosZ*) in soil metagenomes (4). This approach was clearly advantageous compared to the use of an arbitrary e-value threshold by decreasing both the false discovery rate [FDR = FP/(TP + FP)] to about 1% and the false negative rate [FNR=FN/(TP+FN)] to ~ 2%. Accordingly, our approach resulted in a small fraction of false positive metagenomic reads recruited by (or annotated as) reference *nosZ* sequences, i.e., metagenomic reads encoding non-*nosZ* gene fragments but showing a significant score due to the presence of shared domains and/or motifs with *nosZ.* Unlike *nosZ*, other genes sharing highly conserved domains and motifs such as metal binding or ATP-hydrolyzing domains can retrieve a higher fraction of false positive matches when analyzing short-read sequences, therefore representing more challenging cases. Such genes require comparatively higher thresholds in similarity searches in order to achieve a low rate of false positives matches. However, the latter typically comes at the expense of increased frequency of false negatives. Therefore, a variable bitscore threshold across the sequence of the target gene, which would be stringent in highly conserved, non-discriminative regions in order to minimize false positives

but can be lowered in less conserved regions in order to avoid false negatives, should be advantageous compared to the common practice of using arbitrary fixed e-value thresholds. To the best of our knowledge, the idea of a variable threshold across the sequence of a target protein/gene has not yet been implemented in an automated bioinformatic tool.

Here we introduce an automated bioinformatic pipeline, called ROCker, which uses the receiver operating characteristic (ROC) curve to estimate the most-discriminating bitscore thresholds in sliding windows across the sequences of a protein family of interest and evaluates non-discriminative domains shared with unrelated proteins. The pipeline takes as input a list of identifiers for proteins of interest (e.g., beta subunit of RNA polymerase, RpoB) and generates a simulated shotgun dataset using sequenced microbial genomes encoding these proteins (i.e., simulated reads from genomes that encode the reference proteins together with reads from non-target regions of the genome). This dataset of known composition is then used as a training dataset for generating a ROCker profile of most discriminating, position-specific, bitscore values across the target protein alignment, which maximize the recovery of true positive and minimize false positive matches. Therefore, a ROCker profile essentially represents an adaptable filter for minimizing FDR and FNR in similarity search results to accurately detect metagenomic reads related to a single function of interest. We further tested the effectiveness of ROCker with available short-read metagenomes and assessed the diversity of nitrogen cycle genes in terrestrial soils and marine sediments.

2.3 MATERIAL AND METHODS

2.3.1 Implementation

ROCker is implemented in the Ruby programming language and its workflow consists of five tasks. **(i) Build:** Reads a user-provided list of UniProt (Universal Protein Resource) protein identifiers and downloads the corresponding whole genome sequences encoding these proteins for generating datasets that simulate shotgun, short-read, Illumina metagenomes using GRINDER (5). A second list of known negative references, i.e., closely related proteins that should not be considered as true matches can also be given at this step in order to increase the performance of ROCker (see *amoA* example below). The training reference sequences are downloaded and annotated using the European Bioinformatics Institute (EBI) REST API (6) and aligned using ClustalΩ (7). Subsequently, ROCker queries the reference protein sequences provided against the simulated shotgun datasets using BLASTx (8) or DIAMOND (9). **(ii) Compile:** Translates search results to alignment columns, and identifies the most discriminant bitscore per alignment in a 20 amino acid window (or another, user-defined length) in a set of sequences using pROC (10). The latter algorithm calculates sensitivity and specificity using the number of true and false positive matches in each window. The bitscore thresholds are calculated as the value in the ROC curve that maximizes the distance to the identity line (i.e., the non-discriminatory diagonal line in the ROC curve) according to the Youden method. Windows are iteratively refined to reduce low-accuracy regions (<95% estimated accuracy), for all windows with sufficient data (≥5 amino acid positions and ≥3

true positives available). Thresholds in regions with insufficient data are inferred by linear interpolation of surrounding windows. **(iii) Filter:** Uses the calculated set of bitscore thresholds (as estimated by the compile task) to filter the result of a preexisting search. **(iv) Search:** Executes a search of metagenomic sequences against target protein sequences (i.e., single protein function) using BLASTx or DIAMOND, and filters the output according to the most-discriminating bitscores calculated in the Compile step. **(v) Plot:** Generates a graphical representation of the alignment, the thresholds, and the matches obtained, together with summary statistics (Appendix A, Figure A.1).

2.3.2 Target Gene Sequences

Protein sequences for nitrogen cycle reference genes were obtained from the National Center for Biotechnology Information (NCBI) (downloaded in March 2014) and Uniprot (downloaded in June 2015). In order to avoid mis-annotated references, all protein sequences were aligned and visually inspected for the presence of characteristic amino acids or protein motifs and their phylogenetic relationships. Having a list of well-curated reference sequences is key for accurate ROCker results. All reference protein sequences used in the analysis for NirK (n=147), NosZ (n=173), PmoA (n=9), archaeal AmoA (n=5), bacterial AmoA(n=7) and RpoB (n=757) are available through http://enve-omics.gatech.edu.

2.3.3 Simulated Datasets and Benchmark Analyses

2.3.3.1 Generation of Simulated Shotgun Datasets

Simulated datasets were constructed using the “Build” function in ROCker based on an input list of UniProt identifiers for each protein sequence (-P option). GRINDER's parameters differed from their default options as follows: sequencing depth of 3 (for NosZ and NirK, 10 for bacterial and archaea AmoA simulated datasets), remove “-~*NnKkMmRrYySsWwBbVvHhDdXx” characters, sequencing error “uniform 0.1”, mutation ratio “95 5”, and read length distribution “*L* uniform 5”, where *L* is the average read length of the simulated dataset. Simulated datasets ranged from 1 to 43 million reads in size (Appendix A, Table A.1). The CPU time (cput) in hours required for generating simulated datasets can be approximated by using a power law regression as follows: $cput=3.0672*D^{1.096}$ ($r^2=0.948$), where D is the number of protein reference sequences used. Calculated ROCker profiles can be re-used in following similarity searches. The processing of a similarity search output (i.e., ROCker-based filtering) typically takes from a few seconds to a couple minutes on a personal computer, depending on the number of matching sequences.

2.3.3.2 Similarity Search Analysis

The simulated shotgun datasets were used as query sequences for BLASTx (BLAST+2.2.8) and DIAMOND (v0.7.9.58) searches against the reference protein sequences that corresponded to the input UniProt IDs. Default settings were used for BLASTx except that e-value was set to 0.01. For DIAMOND, the settings used were “min score” of 20 and “sensitive”. These settings were used to make DIAMOND comparable to BLASTx in terms of sensitivity, albeit at the expense of speed; users that want faster DIAMOND

searches should opt for the default settings instead. In all cases, only best matches were considered by using the script BlastTab.best_hit_sorted.pl from the enveomics collection (11). The BLASTx searches were used for generating ROCker profiles for NosZ, NirK and RpoB protein references (profiles available through http://enve-omics.ce.gatech.edu/rocker). Hidden Markov models for each set of proteins were built using full-length alignments with HMMer (12). For hidden Markov model (HMM)-based searches, the read sequences were first translated to amino acids using FragGeneScan (13), and subsequently used as query sequences in the hmmsearch algorithm implemented in HMMer (12) (Appendix A, Table A.2).

2.3.4 Tenfold Cross-validation Calculations

Both NosZ and NirK ROCker profiles were further evaluated by performing a tenfold cross-validation test. To ensure that multi-copy references encoded in the same genome were grouped together in cross-validation sets, we randomly separated the genomes into ten subsets (rather than using protein UniProt identifiers). For each subset, a simulated dataset was generated as a query (Test) to challenge a ROCker profile built with the remaining nine subsets (Model). Similarity searches were performed using BLASTx with the parameters described above. FNR and FDR were calculated for each subset and for 100, 150, 200, 250, and 300 bp read length simulated datasets. All generated datasets are available through http://enve-omics.ce.gatech.edu/data/rocker.

2.3.5 Shotgun Metagenomes

Publicly available shotgun metagenomes were downloaded from the Sequence Read Archive (SRA), Metagenomics RAST (MG-RAST), or other web resources (Appendix A, Table A.3). The datasets included two representative Midwest USA agricultural sites (Havana and Urbana, Illinois, USA) (4), two prairie soils that underwent infrared heating for ten years (warming and control; Oklahoma, USA) (14), tropical (Misiones, Argentina) and boreal forests (Alaska, USA) (15), Alaskan permafrost active layer (Alaska, USA) (16), two beach sands (17) and a deep marine sediment (18) related to the Deepwater Horizon oil spill (Florida, USA), human stool (19), and a waste water enrichment sample (20).

2.3.6 Sequence Processing of Shot-gun Metagenomes

SolexaQA (21) was used for quality trimming of raw Illumina metagenomic reads to extract the longest continuous segment with a Phred score ≥ 20. All paired-end or single reads (when only one read was available) longer than 50bp were used for further analysis.

2.3.7 Fraction of Genomes Encoding Nitrogen Cycle Genes

RpoB (RNA polymerase beta subunit) sequences were obtained from reviewed proteins in UniProt/Swiss-Prot. 757 sequences were visually inspected for conservation of functional domains and complete alignment and were used to construct a simulated dataset and ROCker profile (similar options as above for nitrogen cycle genes but using the “--per-genus” option in the building step in order to reduce redundancy caused by sampling individual species with many representative sequences). Short-reads from soil metagenomes were used as

query sequences for independent BLASTx searches (same settings as above) against the NosZ, NirK, AmoA, or RpoB protein references. The ROCker-filtered or e-value-filtered counts were normalized by the median length of the sequences of each protein reference. The fraction of microbial genomes encoding either *nosZ*, *nirK,* or *amoA* (i.e., genome equivalent) was calculated as the ratio of *nirK*, *nosZ*, or *amoA* read counts to *rpoB* read counts using ROCker profiles or e-values.

2.3.8 Phylogenetic Placement of amoA and nosZ Reads

Protein reference sequences for NosZ or Amoa/PmoA were aligned using ClustalΩ (7) with default parameters. The alignment was used to build a phylogenetic tree in RAxML (22) v8.0.19 (LG model). *nosZ*- or *amoA*-reads were extracted from soil metagenomes using ROCker (BLASTx option), and their protein-coding sequences were predicted using FragGeneScan. The latter sequences were added to the NosZ or Amoa/PmoA protein alignment using MAFFT (“addfragments”) (23) and were placed in the corresponding phylogenetic tree using RAxML EPA (24) (-f v option). An *in house* script (“JPlace.to_iToL.rb” available through http://enve-omics.gatech.edu) was used to prepare the visualization of the generated jplace file (25) in iTOL (26).

2.3.9 Availability and Dependencies of ROCker

The ROCker package, documentation, and pre-computed profiles are available through http://enve-omics.ce.gatech.edu/rocker. ROCker is distributed both as a packaged Ruby gem (https://rubygems.org/gems/bio-rocker) and

source code (https://github.com/lmrodriguezr/rocker) under the terms of the Artistic License 2.0. Complete ROCker execution requires the rest-client and json Ruby gems, as well as R (including the pROC package), NCBI-BLAST+ or DIAMOND, GRINDER, and ClustalΩ or MUSCLE (27). In addition, ROCker models can be built online through http://enve-omics.ce.gatech.edu/rocker-build/.

2.4 RESULTS

2.4.1 ROCker Benchmark

We applied ROCker to identify short-reads in simulated datasets of known composition encoding two denitrification genes, namely nitrite reductase (*nirK*) and nitrous oxide reductase (*nosZ)*, and compared the results to other strategies for filtering the output of similarity searches. For this, two manually verified lists of NirK and NosZ protein identifiers were provided to ROCker (as positive references) to generate simulated datasets of known composition resembling short-read metagenomes of different lengths (Figure 2.1 and Appendix A, Table A.1). The datasets were subsequently searched against NirK and NosZ reference sequences to provide the similarity search outputs for comparisons. The coupling of BLASTx with ROCker yielded substantially better performance compared to using fixed e-values, e.g., ~3 and 15 fold-decrease in FDR when compared to the use of a low stringency e-value of 10^{-5} for NosZ and NirK, respectively (100 bp simulated datasets; Figure 2.2 and Appendix A, Table A.2). However, the use of high e-values (i.e., low stringency) provided similar FNR

results to ROCker. In fact, for NirK simulated datasets of longer read lengths, the FNR was slightly lower by ~0.6% to 1.3% when an e-value of 10^{-5} was used compared to ROCker (Figure 2.2). Nevertheless, the high FDR observed for the same searches (at least 24 times higher, on average, compared to ROCker) makes the use of fixed e-values a less accurate approach. In other words, even though using lower e-values (higher stringency, e.g., 10^{-10}) decreased FDR values, this was at the expense of much higher FNR values. In contrast, ROCker's FDR and FNR values were consistently low for all evaluated datasets (Figure 2.2).

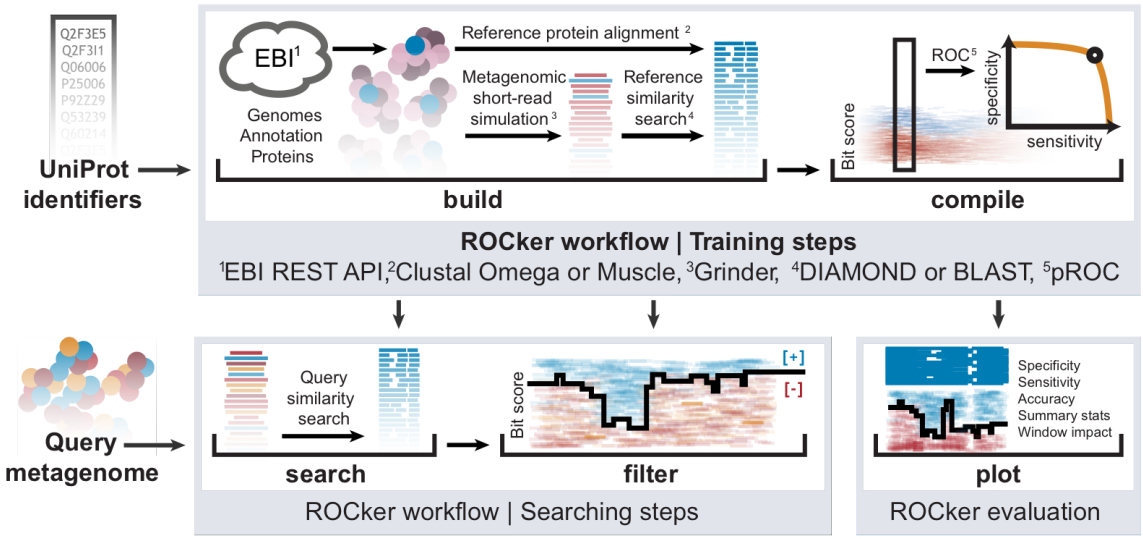


Figure 2.1. ROCker workflow for generating simulated shotgun datasets and calculating position-specific and most-discriminant bitscores.

(Upper panel) ROCker can be used to perform five independent tasks: i) **Build**: Using a user-provided list of unique UniProt protein identifiers for a target protein of interest, ROCker downloads the reference sequences, their corresponding whole genomes and annotation from the European Bioinformatics Institute (EBI) using the REST API. The protein references are aligned and the whole genomes

used for the simulation of short-read Illumina metagenomes and then are searched against the protein reference sequences. The outputs of these searches are then provided to the ii) **Compile function,** where the results are translated to alignment windows where it identifies the most discriminant bitscore that minimizes false positives but maximizes true positive matches. These results are compiled in “ROCker profiles” that essentially represent an adaptable and reusable filter for the output of similarity searches increasing the accuracy of finding a true match compared to the most common practice of using fixed e-value thresholds. (Lower panel) iii) **Search**: Short-read metagenomes are used as query in a similarity search using the target protein sequences as database iv) **Filter**: This tool filters similarity searches using pre-calculated ROCker profiles. Finally, v) **Plot** generates a graphical representation of the ROCker profiles along with the reference sequence alignments and summary statistics (Appendix A, Figure A.1 for an extract of this feature).

In all searches, the recently developed DIAMOND algorithm (using sensitive settings) showed low FNR and FDR when coupled with ROCker, similar to BLASTx (Appendix A, Table and Figure A.2), and was up to ~13-fold faster than BLASTx, consistent with the results reported previously (9). Nonetheless, in every simulation, DIAMOND required more RAM than BLASTx (e.g., 9.6Gb compared to 0.45Gb for the 80bp NirK simulated dataset, respectively). Therefore, the choice of DIAMOND or BLASTx coupled with ROCker would depend on the number of sequences analyzed (e.g., size of metagenomic datasets) and the computational resources available. We also evaluated hidden Markov models (HMM) as implemented in HMMer (12). Searches of both NirK and NosZ simulated datasets showed higher FNR values (about 5-fold higher, on average) compared to ROCker when the same simulated shotgun datasets and reference sequences were used. A better FDR was obtained in HMMer searches compared to the use of a fixed e-value threshold in BLASTx searches, but not as low as those obtained with ROCker (Figure 2.2). Moreover, HMMer required the least amount of memory and was ~860 and 5,700-fold faster, on average, compared to DIAMOND and BLASTx, respectively, consistent with previous results (12). Finally, we compared the results of BLASTx to those of other high-speed protein classification tools such as UproC (28) or GRASP (29) , which showed similar FDR but much higher FNR values (Appendix A, Table A.4). Accordingly, the latter tools were not pursued further.

The evaluation of the performance of ROCker in tenfold cross-validation tests showed low FDR values for both NosZ and NirK ROCker profiles (0.48%

and 1.62%, on average, respectively) in 100, 150, 200, 250 and 300 bp simulated datasets (Appendix A, Figure A.3). However, higher FNR values (5.33% and 17.33%, on average, for NosZ and NirK, respectively) were observed compared to when all references were used for generating ROCker profiles. These results showed that the more reference sequences used when building a ROCker profile and/or the higher the diversity of the reference sequences represented, a better recovery of reads encoding the target gene can be expected. Compared to the use of fixed e-values, ROCker showed lower FDR values in all simulations, consistent with the result reported above. For instance, up to 48 and 35-fold decrease in FDR were observed when compared to the use of low (10^{-5}) and high (10^{-15}) stringency e-values for the NirK simulated datasets, respectively.

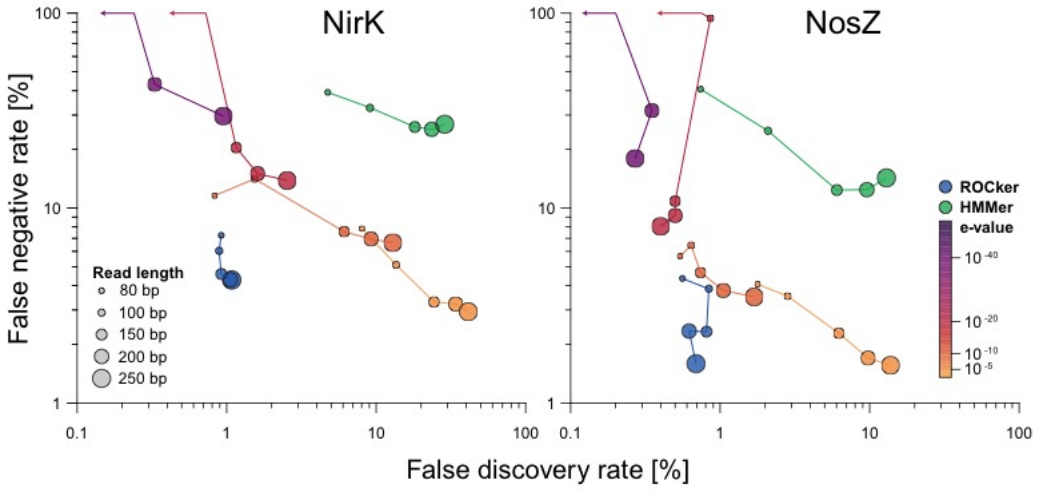


Figure 2.2. Comparison of false negative and false positive rates for simulated shotgun datasets of different read lengths using ROCker profiles and e-value thresholds.

Simulated shotgun datasets of 80, 100, 150, 200, and 250 bp read length (figure legend) were generated using ROCker and searched against reference NirK and NosZ protein sequences using BLASTx. The outputs were filtered using the calculated ROCker profiles (circles in blue) and fixed e-value thresholds (circles in orange to purple gradient). Results from hidden Markov models search of the references NirK and NosZ sequences against the simulated reads are also shown (circles in green).

2.4.2 Targeting a Specific Group of Proteins Using Negative References

It is important to realize that ROCker attempts to optimize the number of matching (simulated) sequences originating from a target gene (true positives) against those originating from the remaining, non-target genes encoded in the same genomes (false positives). If a closely related, yet distinct, protein is encoded by other genomes than those corresponding to the input, simulated sequences from the former genes will not be included in ROCker analyses. To account for such cases and further improve the robustness of the calculated ROCker profile, a second list of non-target, negative references can also be provided to ROCker in order to obtain a filter that can exclude sequences originating from the provided non-target genes, in addition to the other non-target genes encoded in the genomes that correspond to the input. Under this configuration, ROCker simulates datasets generated from both positive (target) and negative references (non-target), and uses them as queries for similarity searches against positive (target) references. However, only matches derived from positive references are considered for determining the position-specific thresholds of the ROCker profile. Using this setup, ROCker was applied to analyze two highly-similar proteins, the bacterial and archaeal ammonia monooxygenase (*amoA*) and the particulate methane monooxygenase (*pmoA*), which are not typically encoded on the same genome and are often challenging to distinguish from each other based on sequence similarity searches. Archaeal AmoA ROCker profiles using bacterial AmoA and PmoA sequences as negative references (Appendix A, Table A.1), showed a moderate decrease of 23-fold and

5-fold in FNR and FDR compared to the use of 10^{-5} and 10^{-10} e-values, respectively (Figure 2.3a). Only low score matches from negative references (considered as false positives) were observed in the similarity search output (Appendix A, Figure A.4), consistent with the higher divergence of archaeal *amoA* from bacterial *amoA* or *pmoA* relative to the divergence between bacterial *amoA* or *pmoA.* In contrast, the performance of the bacterial AmoA ROCker profile using archaeal AmoA and PmoA as negative references was decreased by 66 and 59 fold, on average, for FDR compared to the use of fixed e-values of 10^{-5} and 10^{-10}, respectively (Figure 2.3b). Slightly higher FNR values were observed for bacterial AmoA ROCker profile compared to the archaeal AmoA profile (Figure 2.3b), as expected based on the high sequence similarity between bacterial *amoA* and *pmoA*. The increased FNR values obtained in all searches were attributed to the higher bitscore values calculated for each ROCker profile in order to efficiently discard high-scoring matches derived from negative references (Appendix A, Figure A.4). Therefore, bacterial AmoA ROCker profiles including negative references showed low FDR at the cost of a slightly higher FNR. In summary, having a well-curated set of positive, and, if necessary, negative references is an essential prerequisite for achieving low FDR and FDR values with ROCker.

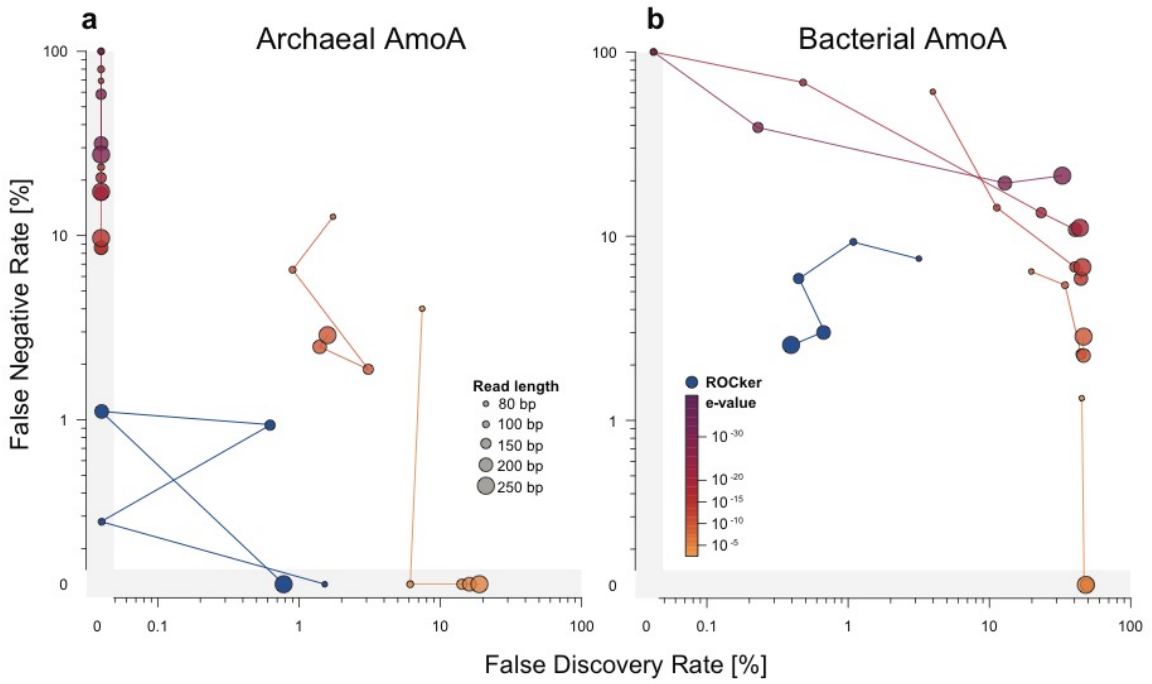


Figure 2.3. Effect of including negative references in AmoA ROCker profiles for simulated shotgun datasets of different read lengths.

Simulated shotgun datasets of 80, 100, 150, 200, and 250 bp read length were searched against (target) AmoA reference sequences using BLASTx. Panel A shows the results of using ROCker archaeal AmoA profiles, including bacterial AmoA and PmoA as negative references, and e-values for filtering the simulated datasets. Panel B shows the results of using ROCker bacterial AmoA profiles, including archaeal AmoA and PmoA as negative references, and e-values.

2.4.3 Using ROCker on Shotgun Metagenomes from Marine and Soil Habitats

2.4.3.1 *nosZ* Gene Abundance in Soil Metagenomes

In order to assess the abundance and diversity of *nosZ* genes in different habitats, we analyzed the phylogenetic classification of *nosZ* gene fragments detected by ROCker (BLASTx search) in ten short-read metagenomes representing agricultural, forest, permafrost and marine sediments (no planktonic samples were analyzed). A maximum likelihood method for the phylogenetic placement of these short reads into a NosZ tree revealed a consistent placement of the recovered fragments according to their habitat of origin (Appendix A, Figure A.5), further supporting that the reads identified by ROCker are indeed NosZ-encoding reads. For instance, the marine genera *Rhodothermus, Maribacter*, and *Caldilinea*, independently recruited ~11 to 320 fold more *nosZ* reads from marine (beach and marine sediments) than terrestrial environments. On the other hand, the *Anaeromyxobacter*, *Opitutus*, and *Gemmatimonas* genera, all commonly found in terrestrial soils, recruited between ~2 and 33 fold more *nosZ* reads from terrestrial than marine environments. The analysis also revealed that atypical or clade II NosZ (4, 30, 31) reads were 2 times more abundant, on average, than the typical or clade I counterparts, which was consistent with our previous analysis using a fixed bitscore threshold across the sequence of NosZ and a smaller set of samples from Midwestern agricultural soils (4). However, typical *nosZ* gene fragments were relatively more abundant in marine sediments than soils, since marine sequences comprised almost 80% of the total typical gene fragments found in all samples.

2.4.3.2 Quantifying *nirK/nosZ* Ratio in Terrestrial and Marine Habitats

The abundance of *nirK* and *nosZ* genes in publicly available short-read metagenomes was quantified based on position-specific bitscore thresholds calculated by ROCker (Figure 2.4a). The use of fixed e-value thresholds (e.g., 10^{-5} or 10^{-10}) generally provided higher abundance estimates compared to those of ROCker, consistent with our expectations from the FDR results reported for simulated datasets. For instance, when a 10^{-5} e-value was used to estimate *nirK* genome equivalents (using universal RpoB protein to normalize abundances), these values exceeded four times, on average, the estimations of ROCker. A similar trend was observed for *nosZ*, albeit ROCker and e-value-based estimates for genome equivalents were closer to each other compared to those calculated for *nirK*, reflecting the less problematic conserved functional domains of NosZ. Further, a higher ratio of *nirK/nosZ* was observed for most terrestrial soil metagenomes compared to metagenomes from sand beaches and sediments when ROCker values were used (Figure 2.4b).

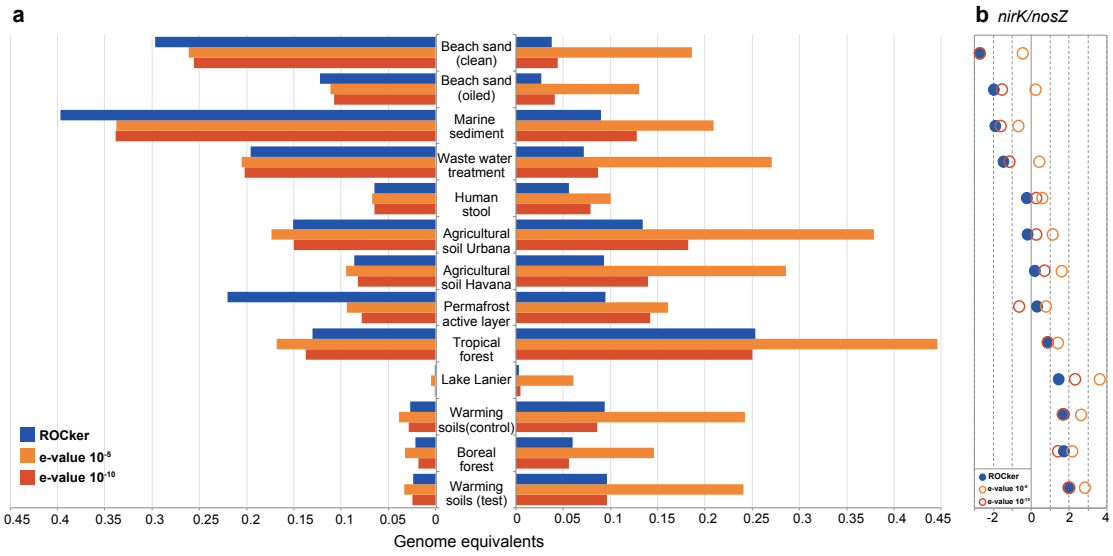


Figure 2.4. Abundance for *nirK* and *nosZ* genes in short-read metagenomes calculated using ROCker or fixed e-value thresholds

Panel A shows the abundance, calculated as the fraction of the microbial community encoding *nirK* or *nosZ*, based on searching short-read metagenomes against NirK (a) and NosZ (b) reference protein sequences. BLASTx searches were filtered using the calculated ROCker profiles or fixed e-values (10^{-5} and 10^{-10}). Panel B shows the log2 ratio of *nirK*/*nosZ* gene abundances using ROCker.

2.4.3.3 Recovering *amoA Gene Fragments from Soil Metagenomes*

We tested the performance of ROCker for extracting bacterial *amoA* reads from soil and sediment shotgun metagenomes (Havana and Urbana soils, and Florida marine sediments) and assessed their phylogenetic placement. Even though more than 30-fold *amoA* reads were extracted when a ROCker profile not including negative references was used (Figure 2.5, inset), only ~10% of these reads were placed in the correct (target) bacterial AmoA clade; the majority of the remaining reads were likely related to PmoA or represented deep-branching members of the membrane-bound monooxygenase (CuMMO) protein family (Figure 2.5b). Conversely, when a bacterial AmoA ROCker profile including negative references (i.e., archaeal AmoA and PmoA) was used to filter the similarity searches, 81% of the *amoA* reads were placed in the expected nodes and branches containing AmoA references (Figure 2.5a).

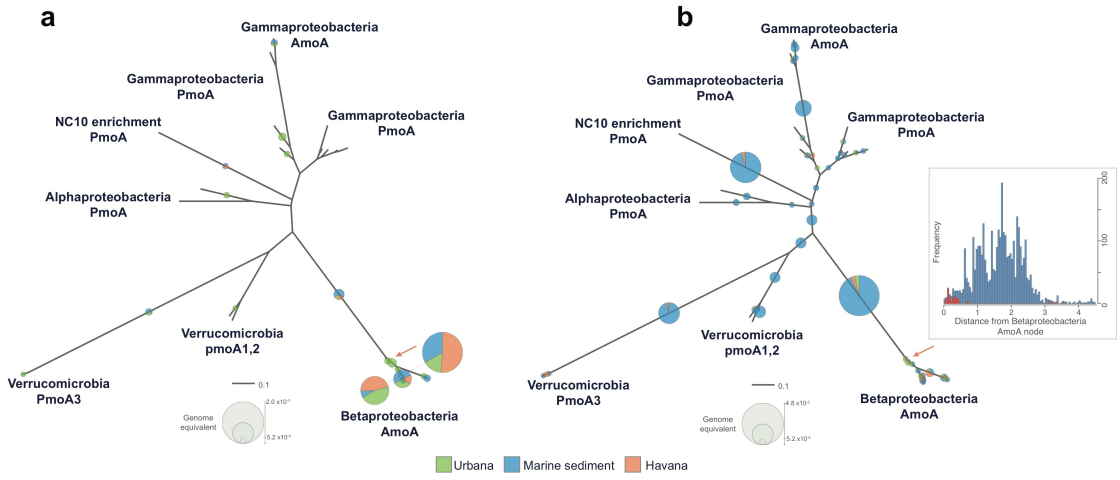


Figure 2.5. Placement of *amoA* reads recovered from terrestrial and marine metagenomes in an AmoA and PmoA phylogenetic tree

A total of 27 bacterial AmoA and PmoA sequences available in the public databases were used to build a reference phylogenetic tree. ROCker bacterial AmoA profiles including (left panel) and not including negative bacterial PmoA references (right panel) were used to identify *amoA* reads from three metagenomes (see figure key). Reads were placed in the phylogenetic tree using RAxML EPA. The radii of the pie charts represent the abundance for each node (calculated as genome equivalents). Note that most reads in the left tree were placed in the betaproteobacterial AmoA clade. However, the reads in the right tree were placed in more deep-branching nodes of the tree or PmoA clades. The inset shows the distribution of the evolutionary distances of the reads from the (target) betaproteobacterial AmoA node (orange arrow), obtained when a ROCker profile including (red bars) and not including (blue bars) negative references was used.

2.4.4 Comparison of ROCker to Alternative Approaches

While several approaches have been recently developed to functionally annotate metagenomic reads (e.g., functional profilers), these tools are based on competitive matches against a large database of functions (28) or they attempt to reconstruct gene variants present in the metagenomes (29, 32), and thus, have different objectives and underlying ideas than ROCker. However, ROCker can be used complementary with these approaches, especially in low sequencing depth metagenomes or with tools that are prone to detect or assemble non-target references (false positives). For instance, in simulated datasets with low sequencing depth for NosZ and NirK (e.g., 1 and 5X), ROCker showed less than 3.33% and 6.6% FNR, respectively, whereas Xander (32) failed to detect and reconstruct more than half of the target sequences (Appendix A, Table A.5). While Xander's performance was better with target sequences showing 10X coverage (e.g., 70-90% of target sequences reconstructed), consistent with results of the earlier study (32), it was still missing target sequences recovered by ROCker (Appendix A, Table A.5 and A.6). Furthermore, in cases where the target references showed high identity to non-related references and also have a different biological role (e.g., AmoA vs. PmoA), ROCker effectively recovered bacterial *amoA*-encoding reads instead of *pmoA* ones (maximum of 3.45% FDR), at the cost of a slightly higher FNR (less than 9.7%, Appendix A, Table A.6). In contrast, Xander showed increased values of FDR (above 30.1%) and FNR (above 10%) due the assembly of false positive non-target references (Appendix A, Table A.5 and A.6). However, when the reads identified by ROCker were

provided as input to Xander, there were no false positive sequences reconstructed by Xander, and Xander's processing time decreased by several orders of magnitude due to the lower sequence complexity of the input. Hence, ROCker can be used complementary to assemblers of target sequences such as Xander in order to increase the accuracy of the reconstructed targets.

2.5 DISCUSSION

The results presented here using ROCker underscore the advantages of using calculated position-specific *versus* fixed thresholds when analyzing short-read metagenomes. E-values depend on the size of the database used and the length of the query sequences, making the determination of the optimal e-value threshold to use a challenging task for short-length queries against different databases. For instance, a closer agreement between ROCker and fixed e-value approaches was observed for NirK abundances in metagenomes when a more stringent 10^{-10} e-value was used (Figure 2.2), but it remains challenging to decide what optimal e-value should be used for other references. In addition, our simulations showed that even considering the bitscore values from the 10% of the best matching reads as thresholds, it is not as robust as ROCker, since such bitscores can represent false positive matches instead. Further, the estimated abundance of proteins with several conserved functional domains such as NirK was frequently overestimated, by at least 2-3 fold, when using fixed e-values (Figure 2.4). Notably, ROCker overcomes these limitations, providing consistent results, independent of the frequency of shared functional domains in the reference of interest.

Two denitrification proteins were chosen to showcase ROCker because they encode a different number of conserved domains, which can increase FDR in similarity searches by recruiting reads encoding similar motifs but originating from non-target (and not related) proteins. NirK is a copper nitrite reductase that contains type-1 and -2 copper centers, commonly found in multicopper oxidases (33). Even though NosZ contains two copper centers, Cu_Z and Cu_A, short-reads of 100 bp or longer have sufficient length in this case to prevent false positive matches from non-*nosZ*-containing reads. Consistent with these characteristics, a 3 to ~5-fold increase in FDR was observed for NirK *versus* NosZ when the e-value strategy (10^{-5}) and different read lengths were used. In contrast, ROCker showed less than 1.5-fold increase in FDR and FNR for NirK *versus* NosZ, for the same datasets (Figure 2.2), consistent with ROCker's ability to robustly deal with genes containing different numbers of conserved domains and/or domains with different degrees of conservation and phylogenetic distribution. Even though low FDR were observed in a tenfold cross validation test, the slightly higher FNR observed was attributable to the reduced sequence diversity in the reference subsets used to generate the ROCker profiles. These findings revealed that users should try to maximize the number of (trusted) reference sequences for building ROCker profiles, and especially the phylogenetic/sequence diversity encompassed by these references for more accurate results. The results presented here for NirK and NosZ illustrate a useful guide for building ROCker profiles and analyzing additional proteins, depending mostly on the number of

conserved domains and motifs encoded by the target protein of interest and their degree of sequence conservation.

It is also important to note that a ROCker profile, while computationally demanding to create (e.g., building *in-silico* datasets) and labor intensive (e.g., manual checking of reference sequences) at the building step (but not for filtering a similarity search output), needs to be built only once and can be subsequently used multiple times, such as in similarity searches for different metagenomic datasets.

We also evaluated popular, alternative algorithms to BLASTx for the similarity search step, including the recently described DIAMOND (9), and hidden Markov models (HMM) as implemented in HMMer (12). ROCker results using DIAMOND (Appendix A, Figure A.2) were faster and comparable in terms of FDR and FNR with BLASTx and thus, the former configuration is recommended for studies with limited computational time available without compromising sensitivity (Appendix A, Table A.2).

ROCker is intended to accurately detect short metagenomic fragments related to a single gene function rather than performing a complete gene functional profile or reconstructing full target sequences from metagenomes. Nonetheless, ROCker can be used complementary to the latter approaches and thus, leads to more accurate analyses of abundance and diversity of target genes in metagenomes. For instance, ROCker showed to be advantageous compared to tools for reconstructing target sequences such as Xander,

especially when the target gene sequences had low sequencing depth (e.g., below 5X), or they were prone to be mistakenly identified as their highly-related but functionally distinct (non-target) gene families (e.g., AmoA vs. PmoA; Appendix A, Table A.5). Having full-length sequences reconstructed from metagenomes enables downstream analyses of the naturally occurring diversity (e.g., diversity surveys, design improved PCR primers); hence, an approach that combines ROCker with tools like Xander could strengthen future studies.

Copper-containing membrane-bound monooxygenase (CuMMO) enzymes catalyze the oxidation of ammonia (AMO), methane (pMMO), and other hydrocarbons, and are encoded in the genomes of methanotrophs and nitrifiers (34-38). Subunit ”A” is typically used as a diagnostic marker of the specific substrate of the enzyme (39). Even though PCR primers can effectively distinguish between bacterial and archaeal *amoA* (40, 41), differences in sensitivity and performance have been identified for primers intended to discriminate between *pmoA* and *amoA* genes (42). These difficulties are mostly due to the high similarity at the nucleotide level because of their recognized evolutionary relatedness (43). To deal with such cases of high sequence identity between target versus non-target genes, especially when the latter are encoded by different genomes than those encoding the former, we implemented the use of negative references for generating ROCker profiles. Remarkably, bacterial AmoA ROCker profiles including PmoA sequences as negative references showed 60-fold improvement in FDR compared to the use of a fixed e-value (e.g., 10^{-5}) (Figure 2.3b), and almost all reads identified were placed in the target bacterial

AmoA tree clade, unlike reads extracted using a ROCker profile without negative references (Figure 2.5, a versus b panels). The use of negative references is also recommended when discrimination between different variants or clades of the same gene family is intended. However, it is important to point out that the decrease in FDR when including negative references was at the expense of a slightly increased FNR, by about 8%, on average, according to our simulated AmoA datasets of different read lengths. Therefore, unless discrimination between closely related protein sequences encoded by the same or different genomes is required, the use of negative sequences should be avoided in order to maximize the number of reads detected that encode the target gene (true positives).

Interestingly, the analysis of soil metagenomes showed a higher ratio of *nirK/nosZ* for terrestrial samples relative to marine sediments (Figure 2.4b), in agreement with previous results based on quantitative real-time PCR (44, 45). These findings are consistent with the hypothesis that in some environments a high fraction of denitrifiers does not possess the genetic potential to reduce N_2O, a potent greenhouse gas. Assuming that gene abundance can be used as a proxy for gene activity (46), these results imply that microbial-mediated reduction of N_2O might be higher (and hence, emissions might be lower) in marine sediments than on land, which remains to be experimentally verified.

Recent studies have shown that previous efforts to determine the abundance of *nosZ* genes have missed a group of divergent sequences, the so-called atypical sequences or clade II, which are functional as N_2O reductases

and are frequently more abundant than their more studied, typical counterparts (4, 30, 31). Consistently, ROCker identified twice as many reads, on average, encoding atypical *versus* typical *nosZ* gene fragments in ten short-read metagenomes representing terrestrial and marine environments. Phylogenetic placement of these short-reads into a NosZ tree revealed that typical *nosZ* reads were mostly derived from marine sediments (Appendix A, Figure A.5), probably reflecting differences in nitrogen cycle pathways and/or regulation between these environments. For instance, typical *nosZ* genes are frequently associated with complete denitrifiers (30), which might account for the higher N_2O reduction potential detected in marine sediments compared to soils. Many atypical *nosZ* reads found in the terrestrial metagenomes were affiliated with the *Anaeromyxobacter*, *Opitutus*, and *Gemmatimonas* genera, and accordingly *nosZ* sequences assigned to these taxa have been frequently recovered from soils based on PCR and/or cloning approaches (30, 47, 48). The high consistency observed between the results of the phylogenetic placement of *nosZ* reads and the habitats of origin of the reads are also in agreement with previous literature and further corroborates the robustness of ROCker.

The only input required to generate simulated datasets and calculate position-specific, most-discriminant bitscores, is a list of UniProt protein sequence identifier numbers for the protein of interest. It should be pointed out, however, that these reference sequences should be carefully selected to represent the protein family of interest (target), as opposed to closely-related homologs of distinct function (when available), in order to obtain accurate

ROCker results. Sequences of related, yet distinct, protein families (negative sequences), which could provide false-positives during similarity searches, can be also given to ROCker in order to increase the performance of the profiles during the “build” stage. Therefore, careful, manual curation of the reference sequences is typically the most time-consuming step of ROCker, and the only step that is not currently fully automated. In our experience, using protein families generated automatically or unsupervised commonly brings error/noise to generated ROCker models, and thus, is not recommended. A few manually curated repositories such as the Functional Gene Pipeline and Repository (FUNGENE) (49) have started to become available, although they are still limited in the number of protein families they encompass.

Finally, finding reads distantly related to the target references might be challenging for ROCker (as is the case for any similarity search-based approach) since ROCker's thresholds (bitscores) are often high, reflecting close similarity to the reference set (particularly in conserved domains present in reference sequences). Using high e-value cutoffs might be advantageous for the latter purpose, albeit at the cost of an unknown (and probably high) number of false positive matches.

In summary, ROCker expands the molecular toolbox for clinical and environmental surveys in the prokaryotic and eukaryotic domain, providing a pipeline to efficiently detect and quantify the abundance of gene fragments of interest in short-read metagenomes. The idea behind ROCker can also be

extended beyond metagenomics to (full-length) protein searches and have broad applications in bioinformatic sequence analysis.

2.6 REFERENCES

1. Kunin,V., Copeland,A., Lapidus,A., Mavromatis, K., and Hugenholtz,P . (2008) A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev.,* **72**, 557–78.

2. Huson,D.H., Auch,A.F., Qi,J. and Schuster,S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.

3. Gerlach,W. and Stoye,J. (2011) Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res.*, **39**, e91.

4. Orellana,L.H., Rodriguez-R,L.M., Higgins,S., Chee-Sanford,J.C., Sanford,R.A., Ritalahti,K.M., Löffler,F.E. and Konstantinidis,K.T. (2014) Detecting nitrous oxide reductase (NosZ) genes in soil metagenomes: method development and implications for the nitrogen cycle. *mBio*, **5**, e01193–14.

5. Angly,F.E., Willner,D., Rohwer,F., Hugenholtz,P. and Tyson,G.W. (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.*, **40**, e94.

6. McWilliam,H., Li,W., Uludag,M., Squizzato,S., Park,Y.M., Buso,N., Cowley,A.P. and Lopez,R. (2013) Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.*, **41**, W597–600.

7. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539–539.

8. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

9. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

10. Robin,X., Turck,N., Hainard,A., Tiberti,N., Lisacek,F., Sanchez,J.-C. and Müller,M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.

11. Rodriguez-R,L.M. and Konstantinidis,K.T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. 10.7287/peerj.preprints.1900v1.

12. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol*, **7**, e1002195.

13. Rho,M., Tang,H. and Ye,Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.*, **38**, e191.

14. Luo,C., Rodriguez-R,L.M., Johnston,E.R., Wu,L., Cheng,L., Xue,K., Tu,Q., Deng,Y., He,Z., Shi,J.Z., *et al.* (2014) Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. *Appl. Environ. Microbiol.*, **80**, 1777–1786.

15. Fierer,N., Leff,J.W., Adams,B.J., Nielsen,U.N., Bates,S.T., Lauber,C.L., Owens,S., Gilbert,J.A., Wall,D.H. and Caporaso,J.G. (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 21390–21395.

16. Mackelprang,R., Waldrop,M.P., DeAngelis,K.M., David,M.M., Chavarria,K.L., Blazewicz,S.J., Rubin,E.M. and Jansson,J.K. (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* **480**, 368–371.

17. Rodriguez-R,L.M., Overholt,W.A., Hagan,C., Huettel,M., Kostka,J.E. and Konstantinidis,K.T. (2015) Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. *The ISME Journal*, **9**:1928–1940.

18. Mason,O.U., Scott,N.M., Gonzalez,A., Robbins-Pianka,A., Bælum,J., Kimbrel,J., Bouskill,N.J., Prestat,E., Borglin,S., Joyner,D.C., *et al.* (2014) Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *The ISME Journal*, **8**, 1464–1475.

19. Consortium,T.H.M.P. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

20. McIlroy,S.J., Albertsen,M., Andresen,E.K., Saunders,A.M., Kristiansen,R., Stokholm-Bjerregaard,M., Nielsen,K.L. and Nielsen,P.H. (2014) 'Candidatus Competibacter'-lineage genomes retrieved from metagenomes reveal functional metabolic diversity. *The ISME Journal*, **8**, 613–624.

21. Cox,M.P., Peterson,D.A. and Biggs,P.J. (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, **11**, 485.

22. Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

23. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, **30**, 772–780.

24. Berger,S.A., Krompass,D. and Stamatakis,A. (2011) Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.*, **60**, 291–302.

25. Matsen,F.A., Hoffman,N.G., Gallagher,A. and Stamatakis,A. (2012) A format for phylogenetic placements. *PLoS ONE*, **7**, e31009.

26. Letunic,I. and Bork,P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, **39**, W475–W478.

27. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

28. Meinicke,P. (2015) UProC: tools for ultra-fast protein domain classification. *Bioinformatics*, **31**, 1382–1388.

29. Zhong,C., Yang,Y. and Yooseph,S. (2015) GRASP: guided reference-based assembly of short peptides. *Nucleic Acids Res.*, **43**, e18.

30. Sanford,R.A., Wagner,D.D., Wu,Q., Chee-Sanford,J.C., Thomas,S.H., Cruz-García,C., Rodríguez,G., Massol-Deyá,A., Krishnani,K.K., Ritalahti,K.M., *et al.* (2012) Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 19709–19714.

31. Jones,C.M., Graf,D.R., Bru,D., Philippot,L. and Hallin,S. (2013) The unaccounted yet abundant nitrous oxide-reducing microbial community: a potential nitrous oxide sink. *The ISME Journal*, **7**, 417–426.

32. Wang,Q., Fish,J.A., Gilman,M., Sun,Y., Brown,C.T., Tiedje,J.M. and Cole,J.R. (2015) Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome*, **3**, 1.

33. MacPherson,I.S. and Murphy,M.E.P. (2007) Type-2 copper-containing enzymes. *Cell. Mol. Life Sci.*, **64**, 2887–2899.

34. Hooper,A.B., Vannelli,T., Bergmann,D.J. and Arciero,D.M. (1997) Enzymology of the oxidation of ammonia to nitrite by bacteria. *Antonie Van Leeuwenhoek*, **71**, 59–67.

35. Lieberman,R.L. and Rosenzweig,A.C. (2005) Crystal structure of a membrane-bound metalloenzyme that catalyses the biological oxidation of methane. *Nature*, **434**, 177–182.

36. Könneke,M., Bernhard,A.E., la Torre,de,J.R., Walker,C.B., Waterbury,J.B. and Stahl,D.A. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*, **437**, 543–546.

37. Tavormina,P.L., Orphan,V.J., Kalyuzhnaya,M.G., Jetten,M.S.M. and Klotz,M.G. (2011) A novel family of functional operons encoding methane/ammonia monooxygenase-related proteins in gammaproteobacterial methanotrophs. *Environ Microbiol Rep*, **3**, 91–100.

38. Lawton,T.J., Ham,J., Sun,T. and Rosenzweig,A.C. (2014) Structural conservation of the B subunit in the ammonia monooxygenase/particulate methane monooxygenase superfamily. *Proteins*, **82**, 2263–2267.

39. Rotthauwe,J.H., Witzel,K.P. and Liesack,W. (1997) The ammonia monooxygenase structural gene *amoA* as a functional marker: molecular fine-scale analysis of natural ammonia-oxidizing populations. *Appl. Environ. Microbiol.*, **63**, 4704–4712.

40. Leininger,S., Urich,T., Schloter,M., Schwark,L., Qi,J., Nicol,G.W., Prosser,J.I., Schuster,S.C. and Schleper,C. (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, **442**, 806–809.

41. Jia,Z. and Conrad,R. (2009) Bacteria rather than Archaea dominate microbial ammonia oxidation in an agricultural soil. *Environmental Microbiology*, **11**, 1658–1671.

42. Junier,P., Kim,O.-S., Molina,V., Limburg,P., Junier,T., Imhoff,J.F. and Witzel,K.-P. (2008) Comparative in silico analysis of PCR primers suited for diagnostics and cloning of ammonia monooxygenase genes from ammonia-oxidizing bacteria. *FEMS Microbiology Ecology*, **64**, 141–152.

43. Holmes,A.J., Costello,A., Lidstrom,M.E. and Murrell,J.C. (1995) Evidence that participate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. *FEMS Microbiology Letters*, **132**, 203–208.

44. Henry,S., Bru,D., Stres,B., Hallet,S. and Philippot,L. (2006) Quantitative detection of the *nosZ* gene, encoding nitrous oxide reductase, and comparison of the abundances of 16S rRNA, *narG, nirK, and nosZ* genes in soils. *Appl. Environ. Microbiol.*, **72**, 5181–5189.

45. Čuhel,J., Šimek,M., Laughlin,R.J., Bru,D., Chèneby,D., Watson,C.J. and Philippot,L. (2010) Insights into the effect of soil pH on N(2)O and N(2) emissions and denitrifier community size and activity. *Appl. Environ. Microbiol.*, **76**, 1870–1878.

46. Petersen,D.G., Blazewicz,S.J., Firestone,M., Herman,D.J., Turetsky,M. and Waldrop,M. (2012) Abundance of microbial genes associated with nitrogen cycling as indices of biogeochemical process rates across a vegetation gradient in Alaska. *Environmental Microbiology*, **14**, 993–1008.

47. Sanford,R.A., Cole,J.R. and Tiedje,J.M. (2002) Characterization and description of *Anaeromyxobacter dehalogenans* gen. nov., sp. nov., an aryl-

halorespiring facultative anaerobic myxobacterium. *Appl. Environ. Microbiol.*, **68**, 893–900.

48. Chin,K.J., Liesack,W. and Janssen,P.H. (2001) Opitutus terrae gen. nov., sp. nov., to accommodate novel strains of the division 'Verrucomicrobia' isolated from rice paddy soil. *Int J Syst Evol Microbiol*, **51**, 1965–1968.

49. Fish,J.A., Chai,B., Wang,Q., Sun,Y., Brown,C.T., Tiedje,J.M. and Cole,J.R. (2013) FunGene: the functional gene pipeline and repository. *Frontiers in Microbiology*, **4**, 291.

CHAPTER 3. DETECTING NITROUS OXIDE REDUCTASE (*NOSZ*) GENES IN SOIL METAGENOMES: METHOD DEVELOPMENT AND IMPLICATIONS FOR THE NITROGEN CYCLE

Reproduced with permission from Orellana, L.H., Rodriguez-R, L.M., Higgins, S., Chee-Sanford, J.C., Sanford, R.A., Ritalahti, K.M., Löffler, F.E., and Konstantinidis, K.T. mBio. 2014 July 1st. Copyright © 2014 American Society for Microbiology

3.1 ABSTRACT

Microbial activities in soils such as (incomplete) denitrification represent major sources of nitrous oxide (N_2O), a potent greenhouse gas. One key enzyme for mitigating N_2O emissions is NosZ, which catalyzes N_2O reduction to N_2. We recently described “atypical” functional NosZ encoded by both denitrifiers and non-denitrifiers, which was missed in previous environmental surveys (Sanford et al., PNAS 109:19709-19714, 2012, doi:10.1073/pnas.1211238109). Here, we analyzed the abundance and diversity of both *nosZ* types in whole-genome shotgun metagenomes from sandy and silty loam agricultural soils that typify the Midwest U.S.A. corn belt. First, different search algorithms and parameters for detecting *nosZ* metagenomic reads were evaluated based on *in silico* generated (mock) metagenomes. Using the derived cut-offs, 71 distinct alleles (95% amino acid identity level) encoding typical or atypical NosZ were detected in both soil types. Remarkably, more than 70% of the total *nosZ* reads in both soils were classified as atypical, emphasizing that prior surveys underestimated *nosZ*

abundance. Approximately 15% of the total *nosZ* reads were taxonomically related to *Anaeromyxobacter*, which was the most abundant genus encoding atypical *nosZ*-type in both soil types. Further analyses revealed that atypical outnumbered typical *nosZ* genes in most publicly available soil metagenomes, underscoring their potential role in mediating N_2O consumption in soils. Therefore, this study provided a bioinformatics strategy to reliably detect target genes in complex short-read metagenomes and suggested that the analysis of both typical and atypical *nosZ* is required to understand and predict N_2O flux in soils.

3.2 IMPORTANCE

Nitrous oxide (N_2O) is a potent greenhouse gas with ozone layer destruction potential. Microbial activities can control both the production and the consumption of N_2O, i.e., conversion to innocuous dinitrogen gas (N_2). Until recently, consumption of N_2O was attributed to bacteria encoding “typical” nitrous oxide reductase (NosZ). However, recent phylogenetic and physiological studies have shown that previously uncharacterized, functional “atypical” NosZ are encoded in genomes of diverse bacterial groups. The present study revealed that atypical *nosZ* genes outnumbered their typical counterparts, highlighting their potential role in N_2O consumption in soils and possibly other environments. These findings advance our understanding of the diversity of microbes and functional genes involved in the nitrogen cycle and provide the means (e.g., gene sequences) to study the effect of climate change to N_2O fluxes.

3.3 INTRODUCTION

In recent years, anthropogenic emissions of greenhouse gases have received increasing attention because of their contribution to global warming (1, 2). Prominent among these gases is nitrous oxide (N_2O) (3), which also contributes to ozone depletion (4, 5). The anthropogenic fixation of dinitrogen (N_2), by means of the Haber-Bosch process, has led to the overuse of synthetic nitrogen-based fertilizers in agriculture (1, 6). As a consequence of the increased nitrogen (N) content of soils, atmospheric N_2O concentrations have risen about 20% relative to preindustrial era levels (2). N_2O emissions are largely the result of bacterial pathways controlling the nitrogen cycle. In particular, N_2O is primarily generated as a product of incomplete classic denitrification (i.e., NO_3^- reduction to N_2O via NO_2^- and NO) and secondarily as a by-product of dissimilatory nitrate reduction to ammonia (DNRA) and oxidation of ammonium to nitrite (nitrification) (7, 8). Besides bacterial activities, abiotic processes and fungal denitrification are also thought to be secondary sources of N_2O (9, 10). Model predictions of N_2O consumption in terrestrial environments primarily focuses on the N_2O to N_2 reduction step, presently attributed to several well-studied classical denitrifiers possessing nitrous oxide reductase (NosZ) (7).

Our previous work has revealed the existence of two phylogenetically distinct NosZ clades, one encompassing the typical Z-type NosZ, which is commonly found in the *Alpha*-, *Beta*- and *Gammaproteobacteria*, and a second clade of atypical NosZ present in diverse organisms representing different phyla. Further analysis of sequenced genomes revealed that most of the typical *nosZ*

are found in bacteria capable of complete denitrification (i.e., encoding all the enzymes for converting NO_3^-/NO_2^- to N_2), whereas atypical *nosZ* genes in bacteria with more diverse N metabolism, including those performing DNRA and missing the NO-generating nitrite reductase genes *nirK* and *nirS* (11, 12). Notably, atypical NosZ have been shown to function as nitrous oxide reductases in several bacteria, such as *Wollinella succinogenes* (13, 14), *Geobacillus thermodenitrificans* (15), the soil isolate *Anaeromyxobacter dehalogenans* (11), and several *Bacillus* sp. isolated from soils (16, 17).

Examination of the potential of microbial communities to reduce N_2O to N_2 has been traditionally performed by evaluating *nosZ* gene and/or transcript presence or abundance by PCR (18, 19). Primers targeting *nosZ* genes, however, were designed according to characterized typical *nosZ* gene sequences, and therefore missed the bulk of divergent atypical genes (11, 12). Furthermore, measured N_2O emissions from soils were frequently lower compared to predictions based on (typical) NosZ transcript abundance and dynamics (20, 21). Therefore, it is likely that atypical NosZ abundance accounts, at least in part, for the discrepancy between predicted and observed N_2O flux.

To circumvent the limitations and explore the total natural diversity of *nosZ* genes in the environment, we analyzed short-read metagenomic datasets from various soils and locations. Even though metagenomics can provide a relatively unbiased, PCR-independent view of the diversity and abundance of individual genes present in a sample, several technical challenges must first be addressed. For instance, in metagenomes of highly diverse microbial communities such as

those obtained from soils, the rates of false positives and false negatives when using similarity searches to detect individual genes in assembled contigs or unassembled short-reads has not been rigorously evaluated, with the probable exception of assessing error rates for the purpose of taxonomic classification, i.e., assigning a sequence to a taxon without necessarily evaluating the potential function and sequence diversity (22, 23). Cut-offs that could minimize the number of false positive matches have not been determined for short-read metagenomes; instead arbitrary, predetermined cut-offs based on e-values (i.e., the likelihood of finding a match by chance) represent the common practice (24).

The objective of the present study was to analyze the diversity and abundance of both typical and atypical *nosZ* genes in soils with contrasting physicochemical properties. To this end, we first developed a strategy based on similarity searches to determine appropriate cut-offs for accurately detecting *nosZ*-encoded metagenomic fragments by analyzing *in silico* metagenomes of known sequence composition. Subsequently, we applied this strategy and derived cut-offs to detect *nosZ* reads in metagenomes from two agricultural soils in the U.S. Midwest that have been subjects of an ongoing multi-year study to assess nitrogen cycling processes, as well as in publicly available metagenomes from various soil ecosystems. Our metagenomic, PCR-independent approach provided a comprehensive and quantitative examination of the diversity and abundance of both typical and atypical *nosZ* genes in soils.

3.4 RESULTS

3.4.1 Evaluating Search Algorithms and Cut-offs for Detecting nosZ Genes in Metagenomes

To determine the best algorithm and parameters for detecting *nosZ* reads in 100 bp long read metagenomes, a reference database of manually verified *nosZ* genes pre-clustered at 95% sequence identity was queried against two such *in silico*-generated datasets, library I (representing the whole genome of 122 NosZ-encoding organisms) and II (representing the whole genome of 1,081 bacteria, including those in library I). From a ROC curve analysis of true and false positive rates, the most appropriate bitscore cut-off values were 107 and 52.2 for the BLASTn and BLASTx searches, respectively. These bitscores provided a sensitivity (fraction of correctly classified positive BLAST matches) of 100% and 98.4% for BLASTn and BLASTx, respectively, and a specificity (fraction of correctly classified negative BLAST matches) of 99.9% for both algorithms. An HMM search resulted in a sensitivity of 46% and specificity of 99.9% (Table 3.1). Additional HMM searches, using models that included more typical and atypical NosZ sequences (built from reference sequences in Appendix B, Table B.1), improved the number of *nosZ* reads retrieved from both *in silico* libraries (~56%). Irrespective of the HMM model employed, a lower fraction of *nosZ* reads was captured when using HMM compared to BLAST searches. Most of the reads missed by the HMM-based approach lacked highly conserved amino acid residues and this accounted for the lower performance of HMM searches, consistent with expectations (HMM models are heavily based on

conserved residues). Therefore, remaining analyses were performed using the BLAST algorithms.

Table 3.1. Comparison of BLASTn, BLASTx and HMMer algorithms for retrieving *nosZ* reads from the in silico Libraries I and II.

Method	*In silico* library	Sensitivity [%]	Specificity [%]
BLASTn	I	100.0	99.9
	II	100.0	99.9
BLASTx	I	98.4	99.9
	II	98.4	99.9
hmmsearch (protein)	I	66.6	99.9
hmmsearch (protein)	II	60.2	99.9

To test the limitations in retrieving metagenomic *nosZ* reads, single typical and atypical representative NosZ protein or nucleotide sequences were independently queried against library II. Although, BLASTn had a similar specificity compared to BLASTx, the latter algorithm was able to capture 735% and 270% more reads (i.e., reads annotated as *nosZ* with a bitscore greater or equal to the calculated cut-off for true positives) of the typical and the atypical references, respectively. Therefore, BLASTx was used in the remaining analyses. Using *Bradyrhizobium japonicum* strain USDA 110 typical NosZ as a single reference, reads derived from 74 out of 127 different alleles encoding NosZ were captured and found to be enriched in closely related *nosZ* sequences to the reference sequence. In contrast, reads for only 32 out of 127 alleles encoding NosZ were captured when *Anaeromyxobacter* sp. strain Fw 109-5 atypical NosZ reference was used in the analysis (Figure 3.1, right panel). This

atypical NosZ reference does not share sequence identity to other target sequences in the 54 to 82% amino acid identity range and thus, the lack of moderately related sequences among the target sequences accounts for the results obtained (Figure 3.1, left panel). More importantly, a linear relationship was observed between the fraction of total reads detected and the level of divergence between the full-length reference and target sequences (Figure 3.1), where 50% or more of the reads were detected when the two sequences shared more than 64 or 68% of sequence identity for typical or atypical NosZ reference sequences, respectively.

Further examination of the reads recruited along the typical NosZ reference (Figure 3.2) showed that the true positive matches (i.e., reads derived from *nosZ* genes with a bitscore greater than 52.2) were evenly distributed along the NosZ reference sequence. The N-terminus of the reference sequence (1-60 amino acid positions) was rarely covered by either true or false positive matches, suggesting that this part of the gene should be avoided when assessing *nosZ* abundance in metagenomes.

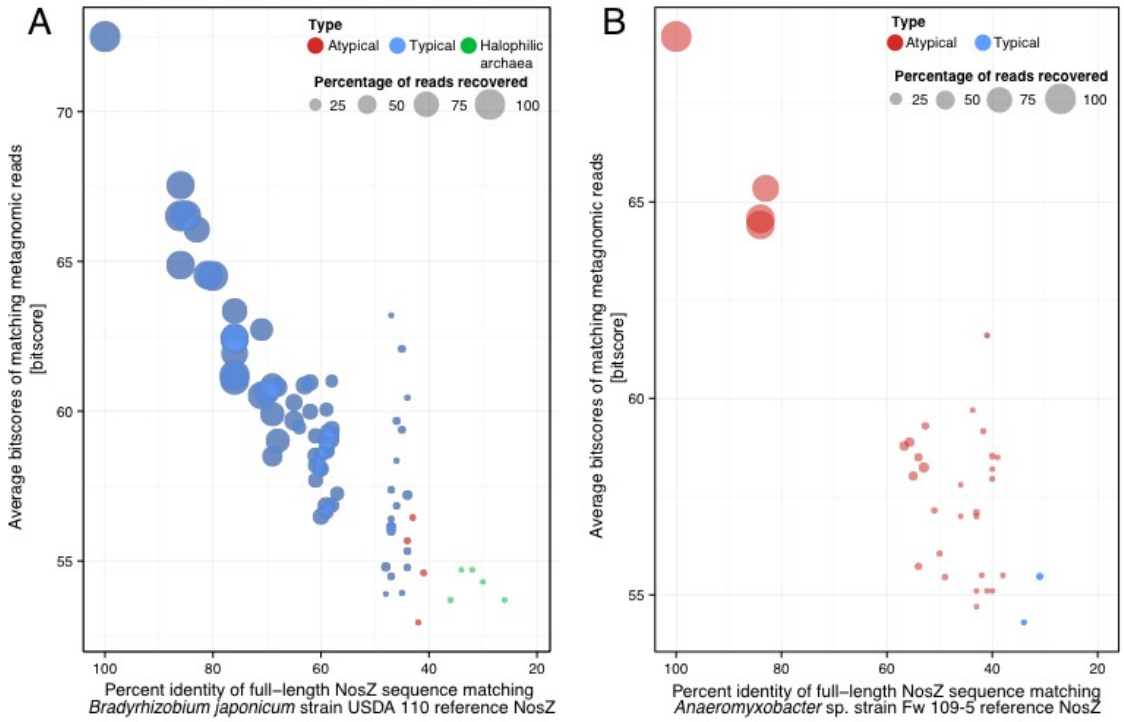


Figure 3.1. Fraction of *nosZ* reads recovered from an *in silico* dataset as a function of their relatedness to the reference query sequence.

nosZ reads were retrieved from the *in silico*-generated library II using *Bradyrhizobium japonicum* strain USDA 110 to represent typical NosZ (left panel) or *Anaeromyxobacter* sp. strain Fw 109-5 to represent atypical NosZ (right panel) reference sequences in a BLASTx search. The average bitscore value (y-axis) for the fraction of *nosZ* reads recovered (circle size) was plotted against the percentage of identity between each reference sequence and the full-length NosZ sequences from which the retrieved (matching) reads originated (target sequence). The linear relationships observed between the fraction of reads detected and the percentage of identity between the full-length target and references sequences were $y=0.464x+40.73$ ($R^2=0.90$) for typical and

$y=0.527x+42.19$ ($R^2=0.89$) for atypical references, where “y” is the percentage of identity between the target and full-length reference and “x” is the fraction of reads retrieved.

3.4.2 Abundance of nosZ Genes in Sandy and Silty Soils

A characterization of the taxa, coverage and assembly statistics of the two soil metagenomes is described in the supplementary results (Appendix B, Supplementary Results B.1). Both soil metagenomes were queried against a 95% identity pre-clustered set of reference NosZ sequences. All matches having a bitscore greater than the calculated cut-off determined based on the *in silico* library analysis were identified as *nosZ* reads and classified as typical or atypical depending on their best match. Atypical *nosZ* reads were clearly the most abundant, comprising 72.9% and 89.6% of the total *nosZ* reads found in the sand and silt loam soil metagenomes, respectively (Figure 3.3). Further, 97% of the *nosZ* reads found in both soil metagenomes (4,929 and 7,280 total in the sandy and silty loam soils, respectively) were recruited by 72 of the 105 NosZ reference sequences, revealing that most of the diversity covered by the references was represented in both soils (Appendix B, Table B.2). In addition, both soil samples showed a similar estimated absolute abundance for *nosZ* reads; ~1.4 $x10^{-5}$ and 2.1 $x10^{-5}$ reads per total reads for Havana sand and Urbana silt loam, respectively (Figure 3.3). The ratio of *nosZ* reads versus single-copy housekeeping gene reads indicated that approximately 16% of the soil bacterial genomes harbored a *nosZ* gene (Appendix B, Table B.3).

Phylogenetic analysis of the atypical *nosZ* reads showed that closed related genes found in members of the *Anaeromyxobacter, Gemmatimonas, Opitutus* and *Hydrogenobacter* genera were the most abundant in both soil samples (Figure 3.4). Additionally, less abundant genera such as *Bradyrhizobium*

and *Rhodopseudomonas*, both known to harbor typical *nosZ* genes, were also present in both soils. Remarkably, atypical *nosZ* reads affiliated with *Anaeromyxobacter* represented 12.7% and 15.2% of the total *nosZ* reads found in both sand and silt loam soils, respectively. The most abundant typical *nosZ* reads were assigned to the *Ralstonia* (3.2%), *Thiobacillus* (3.1%), *Bradyrhizobium* (1.6%) and *Rhodopseudomonas* (1.7%) genera (Appendix B, Table B.2).

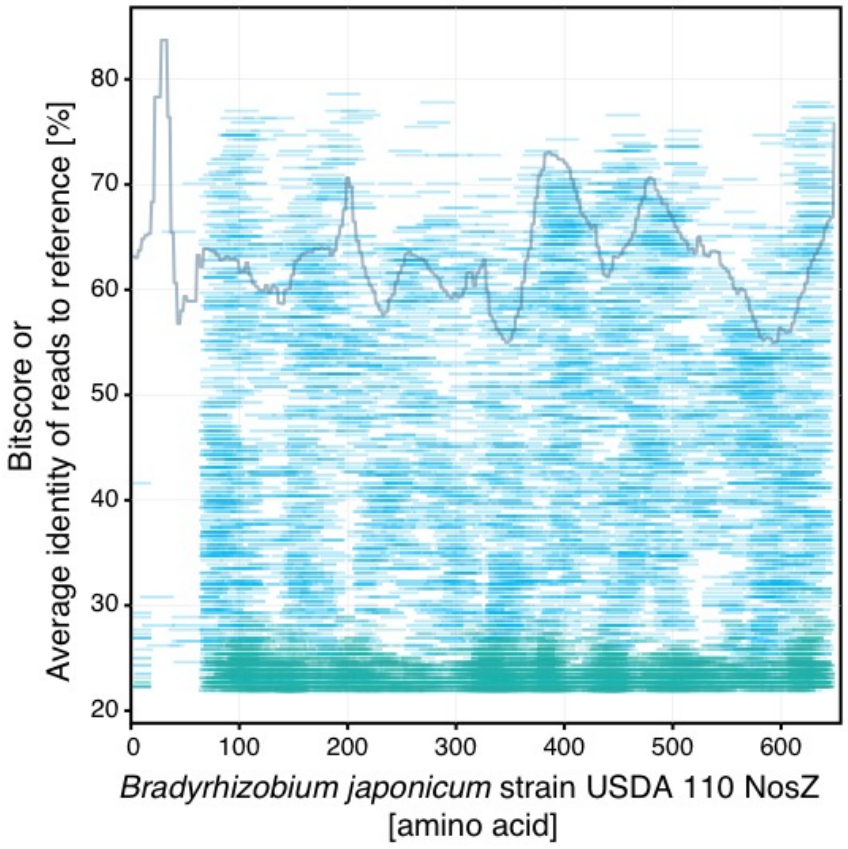


Figure 3.2. Coverage of matching *nosZ* reads from library II along the *Bradyrhizobium japonicum* NosZ reference sequence.

Reads from library II matching the *Bradyrhizobium japonicum* strain USDA 110 NosZ reference are plotted according to their bitscore values. Blue lines represent reads originated from *nosZ* genes and green lines for other genes. The solid line represent the average percent of identity of *nosZ* reads (blue lines) matching the NosZ reference in a three amino acid window.

3.4.3 nosZ Diversity and Abundance in Other Soil Metagenomes

In general, atypical *nosZ* reads were more abundant in the soil metagenomes evaluated (Figure 3.5). The frozen deep-soil permafrost metagenomes [core 1 sample in (25)] showed a greater abundance for typical *nosZ* reads (~80% of total NosZ); however, atypical *nosZ* reads predominated in the upper or active layer (~74% of total *nosZ* sequences). Interestingly, after induced thawing, the microbial communities at both depths showed a small increase in the relative abundance of atypical *nosZ* reads. Furthermore, with the exception of the boreal forest, several biomes studied by Fierer and colleagues (26), including tropical forest, polar and hot desert, arctic tundra, and temperate grassland, showed a higher abundance of atypical vs. typical *nosZ* reads.

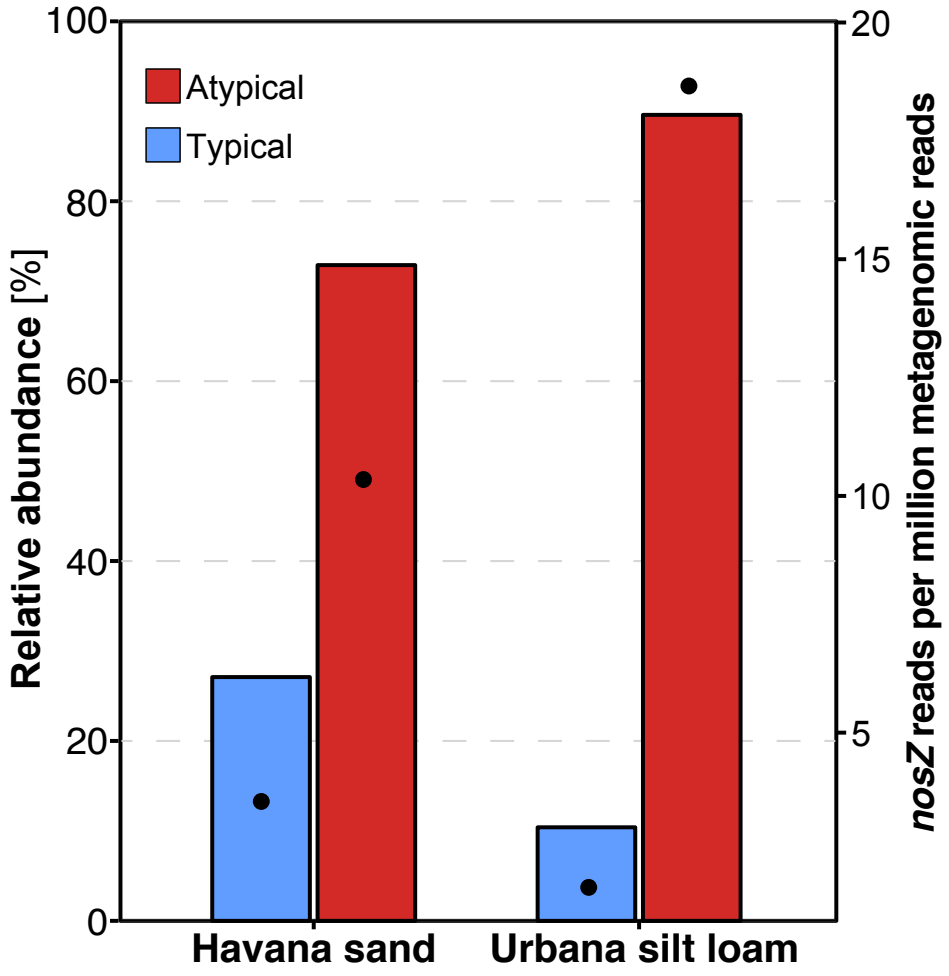


Figure 3.3. Relative abundance for typical and atypical *nosZ* genes in Havana sand and Urbana silt loam soil metagenomes.

All soil metagenomic *nosZ* reads were classified as typical or atypical according to their best match against a reference database of NosZ sequences and the calculated relative abundance of the two gene types are shown (primary y-axis, bars). The absolute abundance, i.e., number of identified *nosZ* reads per million of total reads in each metagenome, is also shown (secondary y-axis, dots).

3.5 DISCUSSION

3.5.1 The Importance of Atypical nosZ

The discovery of functional atypical NosZ has opened the possibility that a much larger number of microorganisms, with previously unaccounted N_2O-reducing potential, could contribute to lessening the N_2O flux into the atmosphere (12). The abundance and diversity of atypical *nosZ* genes were likely missed in previous PCR-based surveys because typical *nosZ* sequences presented the basis for primer design (11, 12) and the two *nosZ* types share only 60.9 +/-8.2% nucleotide identity, on average. In the present PCR-independent metagenome analysis, atypical NosZ sequences were more abundant (>73% of total *nosZ* reads) than their typical counterparts, not only in two agricultural soils differing in physicochemical properties representative of many regions in the Midwest U.S.A., but also in soils from distant geographic locations representing a variety of habitats. Our results were also consistent with the widespread presence of atypical *nosZ* genes, previously hypothesized based on the number of genomes found to encode atypical NosZ among the available genome sequences (11). Therefore, these findings reveal an unexpectedly high potential for N_2O reduction mediated by atypical NosZ in a variety of soil habitats.

It is important to note that our study, being solely based on DNA sequences, evaluated N_2O reduction potential as opposed to the specific *in-situ* activity of NosZ enzymes, typical or atypical. Since negative (purifying) selection efficiently removes unused genes from genomes in microbial populations, the

high abundance of atypical *nosZ* sequences found in different soils samples underpins their functional and/or ecological potential (e.g., Figure 3.5). Given also that N_2O reduction is the only known biochemical function carried by NosZ (11), our results collectively suggest that atypical NosZ are as important, if not more, than their typical counterparts in controlling N_2O fluxes in soils, and likely other environments. Our study also provided the means (e.g., gene sequences for primer design and a bioinformatics strategy) to facilitate future studies of the effect of environmental conditions on NosZ activity and dynamics towards a more predictive understanding of the nitrogen cycle.

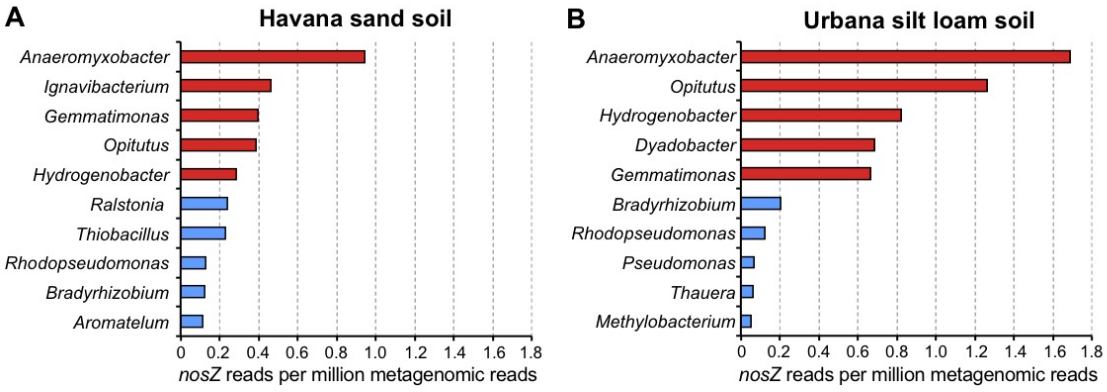


Figure 3.4. Phylogenetic affiliation for the five most abundant genera harboring typical and atypical *nosZ* genes in Havana sand and Urbana silt loam soil metagenomes.

Metagenomic reads were assigned to a genus based on their best match against a reference database of NosZ sequences and the normalized number of reads (based on the size of the datasets) assigned to each genus is shown for Havana sand (A) and Urbana silt loam (B) soils. Red and blue bars represent atypical and typical genes, respectively.

The most abundant *nosZ* genes in the agricultural soils studied here are affiliated with the *Anaeromyxobacter* genus (Figure 3.4). Members of this genus are widely distributed in soils with different physical and chemical characteristics as well as from a variety of geographic locations (11, 27, 28). The high abundance of *nosZ* genes affiliated with members of the non-denitrifying *Anaeromyxobacter* genus is consistent with recent PCR surveys that employed primers targeting atypical *nosZ* sequences (11, 12) and *A. dehalogenans* 16S rRNA gene sequences (11). Further, a high phylogenetic congruence between typical *nosZ* and 16S rRNA gene phylogenies was previously reported (29, 30). Therefore, abundant atypical *nosZ* metagenomic sequences (Figure 3.4 and Appendix B, Table B.2) that have distant matches to homologs of known NosZ-encoding taxa may be harbored by novel taxa. The sequences reported here should facilitate the identification of new taxa, expanding our understanding of phylogenetic diversity on NosZ-encoding soil organisms. The majority of the abundant atypical *nosZ* reads that were assignable to known taxa were found in potentially non-denitrifying genomes of genera such as *Anaeromyxobacter*, *Ignavibacterium*, *Opitutus, Dyadobacter* and *Gemmatimonas*, which were overlooked in previous PCR surveys targeting typical *nosZ* genes. Therefore, the inclusion of these unaccounted N_2O reducers in future environmental studies may help bridge the gap between measured N_2O emissions and denitrification potential based solely on typical *nosZ* genes or *nosZ* transcript measurements.

Our results for the Illinois soils are based on a composite sample comprised of equal DNA quantities extracted from multiple sub-samples,

intended to capture spatial heterogeneity at each field site (at both cm-depth and m-landscape scales). Since the soils were taken at a single time point, it is likely that some of the trends reported here for these soils (e.g., abundance of specific NosZ-encoding taxa, metagenomes average coverage, etc.) might differ temporally, given that agricultural soils typically receive seasonal management inputs. Nonetheless, the high abundance of atypical *nosZ* found in these agricultural soil metagenomes, and in soils from different locations and of diverse physicochemical characteristics (e.g., Figure 3.5), emphasize their potential importance for nitrogen cycling.

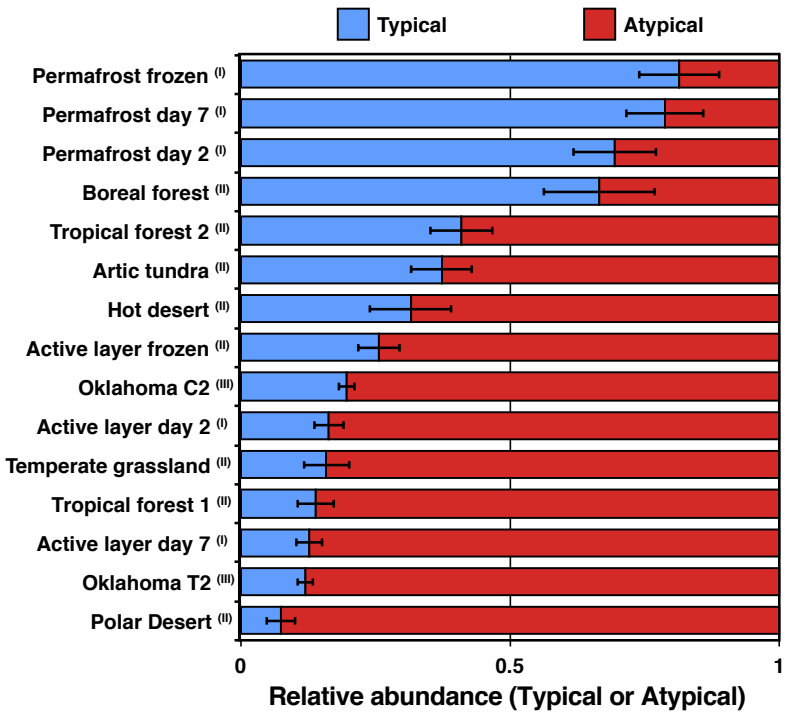


Figure 3.5. Relative abundances of typical and atypical *nosZ* in various soil ecosystems.

nosZ reads were retrieved from available Illumina short-read metagenomes following the same approach used with the Illinois soils reported in this study. The bars represent the probability of finding typical (blue) or atypical (red) as a proxy for their relative abundance and the error bars represent the variance of the sample mean for each soil metagenome. Datasets were obtained from Mackelprang, et al., 2011(I), Fierer, et al., 2012(II) and Luo, et al., 2013 (III).

3.5.2 A Bioinformatics Methodology to Detect Target Genes

Our evaluation of *in silico* generated datasets of known species and gene composition showed that both BLASTn and BLASTx algorithms represent reliable means to detect reads encoding *nosZ* (or other target) genes, albeit with their own strengths and limitations. The selection of the most appropriate algorithm should consider the computational resources available. For example, BLASTx is more computationally demanding than BLASTn, but can capture more divergent sequences if more distantly related sequences/organisms are targeted. However, BLASTn similarity searches are less affected by frameshift-introducing sequencing errors, which might be significant in short-read data even after stringent quality read trimming. Frameshift correction tools such as FrameBot (31), HMM-Frame (32) and FragGeneScan (33) are available to correct these sequencing artifacts and also predict protein coding regions in short reads.

The low performance of the profile-based approach (HMM) versus BLASTx (Table 3.1) is presumably attributable to the lower sensitivity of the former with i) short sequences, ii) sequencing errors creating frameshifts, and iii) reduced fraction of highly conserved amino acid residues specific to the protein of interest, as suggested previously (32, 34). Regardless of these limitations, HMM-based searches are preferable when targeting distantly related homologs and using full-length sequences (e.g., targeting complete gene sequences recovered in assembled contigs).

3.5.3 Recommendations for the Study of Other Genes

The aforementioned approach based on *in silico* metagenomes, the BLAST algorithm, and ROC analysis can be modified for other functional genes of interest. Special attention should be given to conserved domains in the target gene or protein that are shared with other non-target proteins. As shown in Figure 3.2, no false positives matches were observed for the *B. japonicum* strain USDA 110 NosZ reference sequence for bitscores values above the calculated cut-off. The latter finding indicates that no high-identity domains or motifs are shared with other non-NosZ sequences. Other genes may deviate from this pattern and a case-by-case evaluation (e.g., Appendix B, Figure B.1) is recommended. Our approach, when modified to use sliding windows along the sequence of the reference gene, can also determine appropriate cut-offs for different regions of the sequence and identify regions that represent reliable targets for further analyses (e.g., low abundance of false positive matches; PCR primer design).

In silico datasets, simulating different error rates, insert sizes and coverage can be easily constructed to mimic different short-read sequencing technologies or methodologies. Nonetheless, simulating the diversity and variable abundances of individual taxa of real soil metagenomes remains challenging (e.g., our *in silico* datasets had substantially less diversity than the real metagenomes used in the study). The expansion of the *in silico* library I by 13.6 million reads from 959 sequenced genomes not encoding a *nosZ* gene (i.e., *in silico* library II) did not increase the number of false positive matches obtained for *nosZ* (Table 3.1), suggesting that a small number, if any, of false positive

matches should be expected for real metagenomes. In addition, having a comprehensive and well-curated set of protein or gene reference sequences is a key requirement for robust assessment of the best cut-offs and parameters to effectively retrieve reads encoding the gene(s) of interest.

In conclusion, we developed a bioinformatic approach for detection of target genes in short-read metagenomes. This methodology can be extended to the study of any other gene or protein of interest. The high abundance of the previously unaccounted atypical *nosZ* genes in the soil samples suggests that non-denitrifiers and denitrifiers that harbor the atypical *nosZ* may contribute more than previously thought to the reduction of N_2O to innocuous N_2 gas.

3.6 MATERIALS AND METHODS

3.6.1 Samples, DNA Extraction and Sequencing

In November 2011, agricultural soil samples were collected from two sites with long histories of commercial corn and soybean production in the Midwest U.S. corn belt: 1) Havana, IL, (93% sand, 7% clay, latitude 40.296, longitude -89.944, elevation 150 m) and 2) Urbana, IL, (21% sand, 69% silt, 10% clay, latitude 40.075, longitude -88.242, elevation 222 m). In order to provide a metagenome representative of the total soil profile and minimize the effect of sample heterogeneity, soil was collected as three replicate cores (2.5 cm x 30 cm) taken at three locations 30 m apart within each field plot (9 cores total per field), with each core partitioned into four depths (0-5 cm, 5-10 cm, 10-20 cm, 20-30 cm). Soil physicochemical characteristics were measured at each depth (A&L

Laboratories, Ft. Wayne, IN) (Appendix B, Table B.4). DNA was extracted from ~0.5 g of soil from each fraction according to a previously described phenol:chloroform extraction and purification protocol (35) and approximately equal quantities of DNA from each fraction based on agarose gel quantification were pooled to create one composite sample for each soil type. The Illumina Truseq and Nextera DNA library preparation protocols were used for the Havana sand and Urbana silt loam samples, respectively. Sequencing of composite DNA samples was performed using the Illumina HiSeq 2000 platform resulting in 38.4 and 40.2 Gbp of 100 x 100 bp pair-end reads for the Havana sand and Urbana silt loam samples, respectively.

3.6.2 Sequence Processing

An in-house Python script (available at http://enve-omics.gatech.edu) was used for quality trimming of raw Illumina reads as described (36). In brief, this script trims from both the 5' and the 3' end of a sequence using an average Phred score threshold of 20 in 3 bp long windows and discards resulting sequences shorter than 50 bp (Appendix B, Table B.5). The same trimming strategy was applied to publicly available metagenomes for consistency. All BLAST+ (37) and HMMer (38) analyses were based on both single, when the corresponding sister read was not available or discarded after the trimming step, and pair-end reads.

3.6.3 In-silico Libraries and Cut-off Calculation

An in-house Python script was used to generate *in silico* libraries from available complete genomes in the NCBI database as of April 2013 (2,355 sequenced genomes) as described (36). Briefly, this script simulates an Illumina run by generating 100-bp paired-end reads with sequencing error (0.5%), insert size (500 bp), and user-defined coverage (3X). The script also reports the coordinates of the genome from which the reads were generated, so that gene encoding information for each read is available (based on NCBI protein table files or ptt files). The first *in silico* generated dataset (library I) was built based on seven bacterial plasmids and 115 chromosomes previously determined to encode a *nosZ* gene (12). The *in silico* library II was constructed using the 122 DNA sequences from library I and an additional 959 sequenced chromosomes that did not encode NosZ (confirmed independently by BLASTp analysis). Libraries I and II had a total of 7,460 NosZ-encoding reads and were ~14 and ~136 million reads in size, respectively.

BLAST analyses were performed using the BLAST+ 2.2.7 release with the following settings: word size 7, penalty -2, dust no, e-value cut-off 0.001, xdrop gap 150. BLASTx settings were: seg no, e-value cut-off 0.001, word size 3. Previously described *nosZ* nucleotide or protein references from complete genomes were clustered at 95% sequence identity and the longest representative sequence from each cluster was used to construct a reference database, consisting of 54 typical, 47 atypical, and 4 halophilic archaeal representative reference sequences. All reads originating from a *nosZ* gene, whether located on a chromosome or a plasmid, that matched a nucleotide or

protein sequence reference were classified as true positives. Reads not originating from a *nosZ* reference sequence that matched a reference sequence were classified as false positives. The numbers of true and false positives obtained from each algorithm (performance) were evaluated by the receiver operating characteristic curve (ROC) using the R 'pROC' library (39). The bitscore cut-off that maximized performance was calculated as the line that maximized the distance to the identity line (i.e., the non-discriminatory diagonal line where sensitivity and specificity are equal) according to the Youden method for a partial area under the curve (pAUC) between 90% and 100% of specificity (39; see Appendix B, Figure B.1 for a flow chart of the approach).

A Hidden Markov Model (HMM) based on the sequences of six functionally characterized NosZ (*Bradyrhizobium japonicum* USDA 110 27375426, *Wolinella succinogenes* 46934822, *Paracoccus denitrificans* 2833444, *Achromobacter cycloclastes* 37538302 **and *Anaeromyxobacter dehalogenans* 2CP-C** 86158824) was built with HMMer 3.0 and used to query translated reads from Libraries I and II for *nosZ* matches based on the hmmsearch algorithm (38) (Table 3.1). ROC analyses were not performed for HMM searches due to the high specificity and specificity obtained after each search.

3.6.4 Detecting nosZ Reads in Metagenomes

Publicly available metagenomes from Alaskan permafrost (25), soil biomes (26), and soils exposed to a decade of warming (40) were downloaded

from the DOE Joint Genome Institute (www.jgi.doe.gov), MG-RAST (metagenomics.anl.gov), and NCBI Sequence Read Archive webservers, respectively. NosZ-encoding reads (or *nosZ* reads for simplicity) in the above metagenomes were identified by BLAST searches against the 95% identity clustered NosZ reference sequences (Appendix B, Table B.1) and classified as typical or atypical NosZ based on their best match. To account for differences in the number of sequences among the publicly available metagenomes, the presence or absence of each type of *nosZ* read was represented as a binomial distribution for each metagenome. Assuming independence in the presence or absence of each type of *nosZ* gene in each soil sample, the probability of finding either type was calculated from the frequency of *nosZ* reads detected in each metagenome (i.e., a probability closer to one implies a higher abundance for the corresponding type of *nosZ* gene in the metagenome). To account for differences in the number of reads for each metagenome, the standard deviation of the sample mean was calculated for each distribution.

3.6.5 Fraction of Genomes Encoding a nosZ Gene

To estimate the fraction of the microbial populations in the soil community with *nosZ* genes encoded in their genomes the following approach was used. Sequences of three single-copy housekeeping genes (*dnaK*, *recA,* and *rpoB*) were used as references to query each metagenome. The reference set for each housekeeping gene included sequences from 30 different bacterial species (denoted by an asterisk in Appendix B, Table B.1) that also contained a typical or atypical *nosZ* genes (i.e., half of the species in the set harbored a typical *nosZ*

and the other half an atypical gene). The total number of matches obtained in a BLASTn search (settings: no dust, word size 7, penalty -2, max target seqs 1, xdrop 150 and evalue 0.001) for each set of housekeeping and *nosZ* genes was normalized by the average length of the query (reference) sequences. The fraction of the microbial community harboring *nosZ* genes was calculated as the ratio between the normalized number of *nosZ* reads and the reads assigned to each of the housekeeping genes (assuming one *nosZ* gene copy per genome, which is the case for >97% of the analyzed genomes in Appendix B, Table B.1 and Table B.3).

3.6.6 Nucleotide Sequence Accession Number

Both Havana sand and Urbana silt loam metagenomes are available under accession numbers SRR1152189 and SRR1153387 in the Sequence Read Archive server.

3.7 REFERENCES

1. **Canfield DE, Glazer AN, Falkowski PG**. 2010. The evolution and future of Earth's nitrogen cycle. Science **330**:192–6.

2. **Montzka SA, Dlugokencky EJ, Butler JH**. 2011. Non-CO_2 greenhouse gases and climate change. Nature **476**:43–50.

3. **Forster P, Ramaswamy V, Artaxo P, Berntsen T, Betts R, Fahey DW, Haywood J, Lean J, Lowe DC, Myhre G, Nganga J, Prinn R, Raga G, Schultz M, Van Dorland R**. 2007. Changes in atmospheric constituents and in radiative forcing, p129-234. In Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (ed), Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom.

4. **Ravishankara a R, Daniel JS, Portmann RW**. 2009. Nitrous oxide (N_2O): the dominant ozone-depleting substance emitted in the 21st century. Science **326**:123–5.

5. **Portmann RW, Daniel JS, Ravishankara a R**. 2012. Stratospheric ozone depletion due to nitrous oxide: influences of other gases. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **367**:1256–64.

6. **Reay DS, Davidson EA, Smith KA, Smith P, Melillo JM, Dentener F, Crutzen PJ**. 2012. Global agriculture and nitrous oxide emissions. Nat. Clim. Chang. **2**:410–416.

7. **Zumft WG, Kroneck PMH**. 2007. Respiratory transformation of nitrous oxide (N_2O) to dinitrogen by Bacteria and Archaea. Adv. Microb. Physiol. **52**:107–227.

8. **Morales SE, Cosart T, Holben WE**. 2010. Bacterial gene abundances as indicators of greenhouse gas emission in soils. ISME J. **4**:799–808.

9. **Laughlin RJ, Stevens RJ**. 2002. Evidence for Fungal Dominance of Denitrification and Codenitrification in a Grassland Soil. Soil Sci. Soc. Am. J. **66**:1540.

10. **Cooper DC, Picardal FW, Schimmelmann A, Coby AJ, Cooper DC, Picardal FW, Schimmelmann A, Coby AJ**. 2003. Chemical and Biological Interactions during Nitrate and Goethite Reduction by *Shewanella putrefaciens* 200. Appl. Environ. Microbiol. **69***:*3517–3525*.*

11. **Sanford RA, Wagner DD, Wu Q, Chee-Sanford JC, Thomas SH, Cruz-García C, Rodríguez G, Massol-Deyá A, Krishnani KK, Ritalahti KM, Nissen S, Konstantinidis KT, Löffler FE**. 2012. Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils. Proc. Natl. Acad. Sci. U. S. A. **109**:19709–14.

12. **Jones CM, Graf DRH, Bru D, Philippot L, Hallin S**. 2013. The unaccounted yet abundant nitrous oxide-reducing microbial community: a potential nitrous oxide sink. ISME J. **7**:417–26.

13. **Payne WJ, Grant MA, Shapleigh J, Hoffman P**. 1982. Nitrogen oxide reduction in *Wolinella* succinogenes and *Campylobacter* species. J. Bacteriol. **152** :915–918.

14. **Simon J, Einsle O, Kroneck PMH, Zumft WG**. 2004. The unprecedented nos gene cluster of *Wolinella succinogenes* encodes a novel respiratory electron transfer pathway to cytochrome c nitrous oxide reductase. FEBS Lett. **569**:7–12.

15. **Liu X, Gao C, Zhang A, Jin P, Wang L, Feng L**. 2008. The nos gene cluster from gram-positive bacterium *Geobacillus thermodenitrificans* NG80-2 and functional characterization of the recombinant NosZ. FEMS Microbiol. Lett. **289**:46–52.

16. **Jones CM, Welsh A, Throbäck IN, Dörsch P, Bakken LR, Hallin S**. 2011. Phenotypic and genotypic heterogeneity among closely related soil-borne N_2 - and N_2O-producing *Bacillus* isolates harboring the *nosZ* gene. FEMS Microbiol. Ecol. **76**:541–52.

17. **Mania D, Heylen K, van Spanning RJM, Frostegård A**. 2014. The nitrate-ammonifying and *nosZ* carrying bacterium *Bacillus vireti* is a potent source and sink for nitric and nitrous oxides under high nitrate conditions. Environ. Microbiol. doi:10.1111/1462-2920.12478

18. **Scala DJ, Kerkhof LJ**. 1998. Nitrous oxide reductase (*nosZ*) gene-specific PCR primers for detection of denitrifiers and three *nosZ* genes from marine sediments. FEMS Microbiol. Lett. **162**:61–8.

19. **Henry S, Bru D, Stres B, Hallet S, Philippot L**. 2006. Quantitative detection of the *nosZ* gene, encoding nitrous oxide reductase, and comparison of the abundances of 16S rRNA, *narG*, *nirK*, and *nosZ* genes in soils. Appl. Environ. Microbiol. **72**:5181–9.

20. **Cuhel J, Simek M, Laughlin RJ, Bru D, Chèneby D, Watson CJ, Philippot L**. 2010. Insights into the effect of soil pH on N(2)O and N(2)

emissions and denitrifier community size and activity. Appl. Environ. Microbiol. **76**:1870–8.

21. **Henderson SL, Dandie CE, Patten CL, Zebarth BJ, Burton DL, Trevors JT, Goyer C**. 2010. Changes in denitrifier abundance, denitrification gene mRNA levels, nitrous oxide emissions, and denitrification in anoxic soil microcosms amended with glucose and plant residues. Appl. Environ. Microbiol. **76**:2155–64.

22. **Huson DH, Auch AF, Qi J, Schuster SC**. 2007. MEGAN analysis of metagenomic data. Genome Res. **17**:377–86.

23. **Gerlach W, Stoye J**. 2011. Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic Acids Res. **39**:e91. doi: 10.1093/nar/gkr225

24. **Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P**. 2008. A bioinformatician's guide to metagenomics. Microbiol. Mol. Biol. Rev. **72**:557–78.

25. **Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM, Jansson JK**. 2011. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature **480**:368–71.

26. **Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert J a, Wall DH, Caporaso JG**. 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc. Natl. Acad. Sci. U. S. A. **109**:21390–5.

27. **Sanford RA, Cole JR, Tiedje JM**. 2002. Characterization and Description of *Anaeromyxobacter dehalogenans* gen. nov., sp. nov., an Aryl-Halorespiring Facultative Anaerobic Myxobacterium. Appl. Environ. Microbiol. **68**:893–900.

28. **Petrie L, North NN, Dollhopf SL, Balkwill DL, Kostka JE**. 2003. Enumeration and Characterization of Iron(III)-Reducing Microbial Communities from Acidic Subsurface Sediments Contaminated with Uranium(VI). Appl. Environ. Microbiol. **69**:7467–7479.

29. **Jones CM, Stres B, Rosenquist M, Hallin S**. 2008. Phylogenetic analysis of nitrite, nitric oxide, and nitrous oxide respiratory enzymes reveal a complex evolutionary history for denitrification. Mol. Biol. Evol. **25**:1955–66.

30. **Palmer K, Drake HL, Horn MA**. 2009. Genome-derived criteria for assigning environmental *narG* and *nosZ* sequences to operational taxonomic units of nitrate reducers. Appl. Environ. Microbiol. **75**:5170–4.

31. **Wang Q, Quensen JF, Fish JA, Lee TK, Sun Y, Tiedje JM, Cole JR**. 2013. Ecological patterns of *nifH* genes in four terrestrial climatic zones explored with targeted metagenomics using FrameBot, a new informatics tool. MBio **4**:e00592–13. doi:10.1128/mBio.00592-13.

32. **Zhang Y, Sun Y**. 2011. HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors. BMC Bioinformatics **12**:198. doi:10.1186/1471-2105-12-198

33. **Rho M, Tang H, Ye Y**. 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. **38**:e191. doi: 10.1093/nar/gkq747

34. **Zhang Y, Sun Y**. 2012. MetaDomain: a profile HMM-based protein domain classification tool for short sequences. Pac. Sym. Biocomput 271–282.

35. **Welsh A, Chee-Sanford J, Connor L, Löffler F, Sanford R**. 2014. Refined NrfA phylogeny improves PCR-based *nrfA* gene detection. Appl. Environ. Microbiol. doi: 10.1128/AEM.03443-13.

36. **Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT**. 2012. Individual genome assembly from complex community short-read metagenomic datasets. ISME J. **6**:898–901.

37. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL**. 2009. BLAST+: architecture and applications. BMC Bioinformatics **10**:421. doi:10.1186/1471-2105-10-421

38. **Eddy SR**. 2011. Accelerated Profile HMM Searches. PLoS Comput. Biol. **7**:e1002195. doi: 10.1371/journal.pcbi.1002195

39. **Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M**. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics **12**:77. doi:10.1186/1471-2105-12-77

40. **Luo C, Rodriguez-R LM, Johnston ER, Wu L, Cheng L, Xue K, Tu Q, Deng Y, He Z, Shi JZ, Yuan MM, Sherry R a., Li D, Luo Y, Schuur E a. G, Chain P, Tiedje JM, Zhou J, Konstantinidis KT**. 2013. Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. Appl. Environ. Microbiol. **80**:1777-86

CHAPTER 4. YEAR-ROUND METAGENOMES REVEAL STABLE MICROBIAL COMMUNITITES IN AGRICULTURAL SOILS AND NOVEL AMMONIA OXIDIZERS RESPONDING TO FERTILIZATION

Reproduced in part with permission from Luis H. Orellana, Joanne C. Chee-Sanford, Robert A. Sanford, Frank E. Löffler, and Konstantinos T. Konstantinidis. Year-round metagenomes reveal stable microbial communities in agricultural soils and novel ammonia oxidizers responding to fertilization. All copyright interests will be exclusively transferred to the publisher upon acceptance.

4.1 ABSTRACT

The dynamics of microbial communities in agricultural soils, especially after major activities such as nitrogen fertilization, remain elusive but are important for nutrient cycling research. Here, we analyzed 20 short-read metagenomes collected at 4 time points across 1 year from two depths (0-5 and 20-30 cm) in two Midwestern agricultural sites representing contrasting soil textures (sandy versus silty-loam), with similar cropping histories. While microbial community taxonomical and functional differed between the two locations and across soil depths, these diversity components were remarkably stable throughout the year compared to other natural ecosystems. For example, among the 69 population genomes assembled from the metagenomes, 75% showed less than 2-fold change in abundance between any two sampling points. Interestingly, six deep-branching, novel *Thaumarchaeota* and three Comammox nitrifier *Nitrospira* increased up to 5-fold in abundance upon the addition of nitrogen fertilizer at the sandy site. These results revealed that indigenous archaeal ammonia oxidizers

may respond faster (r-strategists) to N-fertilization than previously thought. None of 29 recovered putative denitrifier genomes encoded the complete denitrification pathway, suggesting that denitrification is carried out by a collection of different populations. Altogether, our study identified novel microbial populations and genes responding to seasonal and human-induced perturbations in agricultural soils.

4.2 INTRODUCTION

Agricultural soils are characterized by a dynamic interplay between complex biotic and abiotic processes driving the nutrient cycling of the soil ecosystem (Dick, 1992; Altieri, 1999). Even though the central role of microorganisms participating in the cycling of nutrients in soil ecosystems has been extensively reported (Kennedy and Smith, 1995; Whitman *et al.*, 1998), little is known about the natural population diversity responding to environmental and seasonal agricultural perturbations (e.g., nitrogen fertilization). This scarcity of information limits our understanding of the role of microorganisms generating and consuming key nutrients such as carbon (C) and nitrogen (N), in soils. Rates of synthetic N fertilizer addition often exceed crop requirements, resulting in unintended losses of N-compounds from soil. For instance, higher atmospheric concentrations of nitrous oxide (N_2O), a potent greenhouse and ozone-depleting gas in the stratosphere (1), have been recorded compared to preindustrial-era levels, with soils contributing approximately 65% of the total N_2O emitted to the atmosphere (2). These elevated N_2O emissions result mainly from the activities of microorganisms controlling the N-cycle (3, 4). Therefore, the identification and

tracking of natural microbial communities responding to elevated N inputs can provide important new insights toward better modeling of N_2 and N_2O dynamics in soils, with implications for greenhouse gas emissions.

The advent of high-throughput sequencing technologies has expanded our understanding of natural microbial communities by unraveling previously undetected microbial diversity in different ecosystems. For instance, metagenomic approaches have been applied to examine highly diverse microbial ecosystems, such as soils and sediments, to survey genetic markers related to wide-ecosystem processes, e.g., C and N-cycling genes (5), and to recover microbial genomes previously elusive to culturing approaches (6). Despite these efforts, little is known about the diversity and dynamics of microbial communities undergoing recurrent external perturbations, such as agricultural practices on soils ecosystems. For instance, the impact of agricultural activities on the dynamics of microbial communities involved in the cycling of C (e.g., biomass degradation) and N (e.g., nitrification) during the growing season remains essentially unknown. In addition, the impact of long-term synthetic N fertilization or natural nitrogen fixation on natural microbial communities is crucial for better predicting the cycling and fate of N species in agricultural soils.

Diverse microbial populations catalyze the transformation of soil N by performing complete or incomplete denitrification ($NO_3^-/NO_2^- \rightarrow N_2O$ and/or N_2), nitrification ($NH_4^+ \rightarrow NO_2^-/ NO_3^-$), and ammonification (e.g., protein degradation and respiratory ammonification, NO_3^- or $NO_2^- \rightarrow NH_4^+$). Further, nitrification was

originally described as a sequential two-step process mediated by ammonia-oxidizing bacteria (AOB) and nitrite-oxidizing bacteria (NOB) (Prosser, 1990). However, within the past decade, new genomic and cultivation advances have rapidly broadened our understanding of the microbial diversity participating in nitrification. For instance, studies of ammonia-oxidizing archaea (AOA) belonging to the *Thaumarchaeota* phylum have revealed differences in ecophysiology and substrate affinity for these organisms compared to AOB (7). While AOA numerically dominate over AOB in many soils, the contribution of the former to terrestrial nitrification and consequent N_2O production is not yet well established (8). Moreover, the recent discovery of "Comammox" bacteria related to *Nitrospira* encoding all necessary enzymes to perform complete nitrification in biofilms (9) and bioreactors (10), has initiated new interest into the relative contributions of these organisms to nitrogen cycling in nature. Since ammonia oxidation provides the main source of energy for these microbial groups, it has been proposed that a tight interplay among affinity, tolerance and source of ammonia could control these microbial populations in nature (11, 12). Nevertheless, the relative response of AOA, AOB, NOB and the recently described Comammox populations to agricultural fertilization is currently unclear, but important for predicting their relative contributions to the N-cycle and generation of by-products such as N_2O.

In the present work, we describe the diversity and seasonal dynamics of natural microbial communities in two US Midwest soil ecosystems with contrasting soil textures, i.e., sandy (93% sand; 7% clay) and silty-loam (21%

sand; 69% silt; 10% clay). We analyzed total microbial community DNA (Illumina short-read metagenomes) from samples collected during the four seasons in 2012. Our findings showed that the microbial communities in these agricultural soils are generally stable throughout the year, meaning that they are not characterized by strong seasonal shifts in diversity that typify other natural ecosystems such as freshwater lakes and the ocean. In addition, novel deep branching ammonia-oxidizing *Nitrospirae* and *Thaumarchaeota* are among the most abundant nitrifier organisms in these agricultural soils, especially in the deeper soil layer, and are responsive to nitrogen fertilization. These findings highlight that current understanding of nitrification is based on a limited fraction of the extant nitrifier diversity, and the populations identified by our study can be used as relevant models for studying soil nitrification.

4.3 RESULTS

4.3.1 Agricultural Soil Physicochemical Characteristics and Statistics of Metagenomes

We focused our study on two sites with established legacies from at least 15 years of the same agricultural management practices. These sites represent distinctive soil textures with opposing drainage characteristics and water holding capacities located in Havana (sandy) and Urbana (silt loam), Illinois, USA. Soil cores obtained during 2012 revealed contrasting soil chemical parameters between sites and also depths (Appendix C, Table C.1). In particular, higher organic matter (OM) content was observed in Urbana (average=3.84%)

compared to Havana (average=0.7%) (two-tailed *t*-test, $P < 0.01$). Whereas no statistically significant differences in OM were observed between soil layers in Urbana, a higher OM content was observed for the top soil layer in Havana across the year (two-tailed *t*-test, $P < 0.01$). Even though pH values were higher in Havana (average=7.42) compared to Urbana (average=6.03) (two-tailed *t*-test, $P < 0.01$), contrasting values were also observed within the soil layers in each site. While the 0-5 cm soil layer in Havana was slightly more alkaline (average=7.57) than the 20-30 cm layer (average=7.32) (two-tailed *t*-test, $P=0.057$*)* a higher difference between top (average=5.7) and deep (average=6.26) soil layers was determined in Urbana (two-tailed *t*-test, $P < 0.01$) (Appendix C, Table C.1). During the growing season, Havana was planted with maize and received fertilizer (UAN28; ammonium nitrate 40%, urea 30%, water 30%) and herbicide (glyphosate), whereas Urbana was planted with soybean, received herbicide but did not receive synthetic fertilizer (Appendix C, Table C.2).

20 shotgun metagenomes were obtained from both agricultural sites ranging from 21 to 72 million reads per sample (or ~2.7 to 9.2 Gbp, 4.3 Gpb average) and average read length of ~125 bp after trimming (Appendix C, Table C.3). The estimated coverage based on the read redundancy value calculated by the Nonpareil algorithm (13) revealed an average coverage of 0.15 and 0.14 for the metagenomes obtained from the 0-5 cm soil layers in Havana and Urbana, respectively. Higher coverage values of 0.21 and 0.26 were observed in the metagenomes obtained from the 20-30 cm soil layers in Havana and Urbana, respectively (Appendix C, Table C.3). Sequence diversity values (a measure of

alpha diversity derived from Nonpareil curves), showed higher values for the Havana 0-5 cm soil layer than in the 20-30 cm soil layer, but no differences between Urbana soil layers (two-sided *t*-test, Appendix C, Figure C.1). The co-assembly of the metagenomes by depth and location allowed for the recovery of over 1.1 million contigs over 500 bp in length each, comprising 1.59 Gbp, in total, for the four co-assemblies and ~2.3 million predicted genes (Appendix C, Table C.4). The N50 values averaged from both 20-30 cm soil metagenomes were slightly higher than the 0-5 cm samples (~1,240 vs. 1,650 bp), reflecting the lower sequence coverage determined by Nonpareil for the surface samples.

4.3.2 Microbial Community Structure and Diversity

Throughout the year, stable abundances of functional genes were observed for both soil depths. For instance, the ~60 most abundant functional categories associated to secondary metabolism showed, on average, 5.3% and 6.8% annual variation (measured as coefficient of variation across all samples) for the top and deep layers of Havana, respectively. Similarly, 6.9% and 10.9% average variation for top and deep layers was observed for Urbana, respectively (Figure 4.1). A much lower variation was observed when housekeeping genes and general metabolic functional categories included to the latter analysis (Appendix C, Figure C.2a), as expected for core functions encoded by almost every organism. To assess the extent of within-site variation, 7.2% and 14.6% variation in gene annotations related to secondary metabolism was observed among metagenomes obtained from three independent soil cores (replicates) from the 20-30 cm soil layer in June for Havana and Urbana, respectively

(Appendix C, Figure C.2b). A smaller average variation was observed within Havana (2.01%) and Urbana (5.3%) cores when all functional categories were considered. Thus, these consistent variation patterns between and within sites revealed a higher variation within Urbana soil metagenomes but stable variation values across the year for both sites.

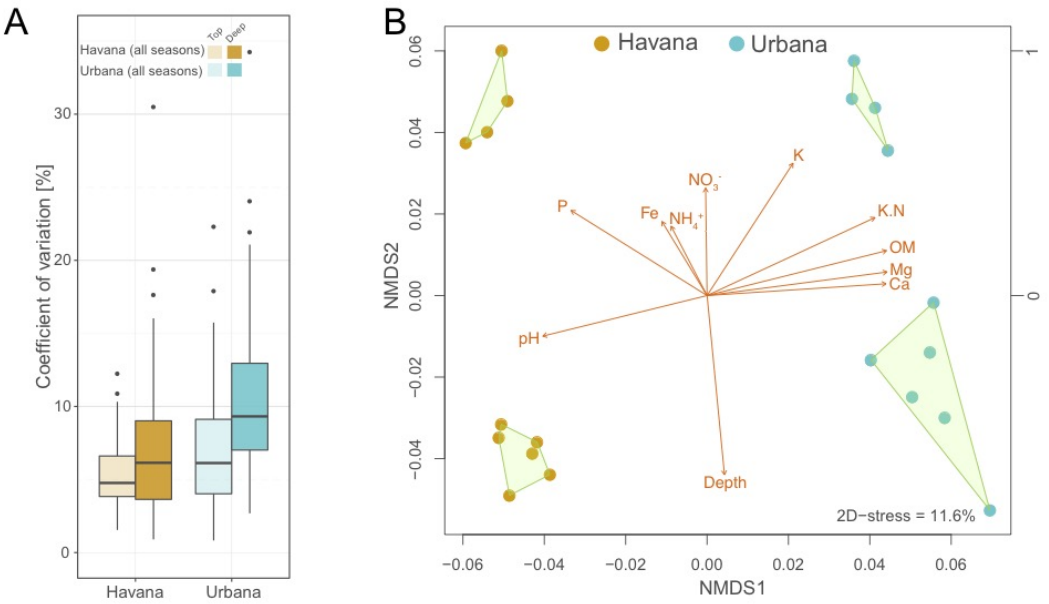


Figure 4.1. Sequence and functional compositional differences between two agricultural sites.

Distributions of coefficients of variation of SEED subsystems related to secondary metabolism in Havana and Urbana. **B.** Non-metric multidimensional scaling (NMDS) analysis based on MinHash distances determined by MASH showed independent clustering by site and depth. The length of the gold arrow is proportional to the correlation between measured metadata and determined ordination values. The direction of the arrow points to increasing changes in the values of the corresponding metadata.

Interestingly, a deep vs. surface separation of samples was observed between metagenomes from each site based on annotation-independent MinHash similarity distances (NMDS, MASH similarity distances and ANOSIM *P*-value ≤ 0.001, R=0.95, Figure 4.1b). A similar spatial clustering was observed for the soil samples when annotations derived from short-reads (SEED subsystems) were used for ordination (NMDS, Bray-Curtis distances and ANOSIM *P*-value ≤ 0.001, R=0.89; Appendix C, Figure C.3a). We further investigated the functional features driving the separation between the soil layers in each site using the SEED subsystem information. Genes encoding oxygen and light sensors, ferrous iron transporters and photolyases, among others, were characteristic of the top layers in both soils (Log_2 fold change >2, p-adjusted <0.05, Appendix C, Figure C.4). Interestingly, 39.4% and 34.1% of these overrepresented SEED subsystems were related to archaeal pathways in the deep soil layers in Havana and Urbana, respectively (Log_2 fold change >2, p-adjusted <0.05, Appendix C, Figure C.4). Significantly higher functional diversity (Chao-Shen entropy index) was observed for Havana compared to Urbana (two-tailed *t*-test, *P*-value < 0.05,) over the entire year of sampling. However, similar diversity values were observed among top and deep soil samples from each site (Appendix C, Figure C.1c).

Similar to the functional annotations, the taxonomic affiliation of recovered 16S rRNA gene fragments from the soil metagenomes showed moderate changes in abundance across the year but significant differences were detected between the soil depths (for further details, see Appendix C, Figure C.3b and Supporting Information C1). We also explored the impact of the microbial

communities on the breakdown and recycling of plant biomass in soils. Similar to previous results, stable abundances for genes associated with biomass and polysaccharide degradation (i.e., glycoside hydrolases) were found throughout the year (details in Supporting information).

4.3.3 Effect of Fertilization on Nitrogen Cycle Gene Abundances

We further examined the impact of agricultural practices on target N cycling genes throughout the growing season. We combined an accurate approach for detecting specific genes in metagenomes (ROCker) with a phylogenic-based classification of detected short-reads (Appendix C, Table C.5). Havana and Urbana showed 4.3-fold and 3.2-fold higher abundance, on average, of archaeal ammonia monooxygenase (*amoA*) and bacterial hydroxylamine oxidase (*hao*), and nitrite oxidoreductase (*nxrA*) genes in the deep layer of soil relative to the top layer, even though both sites received different sources of N (i.e., synthetic N vs naturally fixed N) (Figure 4.2a). Also, urease genes (*ureC)* had the highest relative abundance (~0.3 genome equivalent) among all N genes detected in both sites, but did not show substantial differences in abundance between soil layers (Appendix C, Figure C.5).

As opposed to Urbana, Havana received synthetic N fertilizer during late April 2012, which we hypothesized affected the abundance and distribution of N-cycle genes associated with oxidation processes (e.g., nitrification) such as *amoA*, *hao* and *nxrA* (Figure 4.2a). Interestingly, ammonia oxidation genes showed the highest increase in abundance in June (20-30 cm soil layer), about

one month after fertilization was performed in Havana (3.8 and 1.5-fold increase from April to July for archaea and bacteria). In addition, archaeal *amoA* gene fragments were ~6 times more abundant than bacteria, on average, during the year. On the other hand, Urbana, which relied on naturally fixed N, showed slightly higher abundance for nitrification genes during April, but in general these relative abundances were much more stable compared to Havana throughout the year.

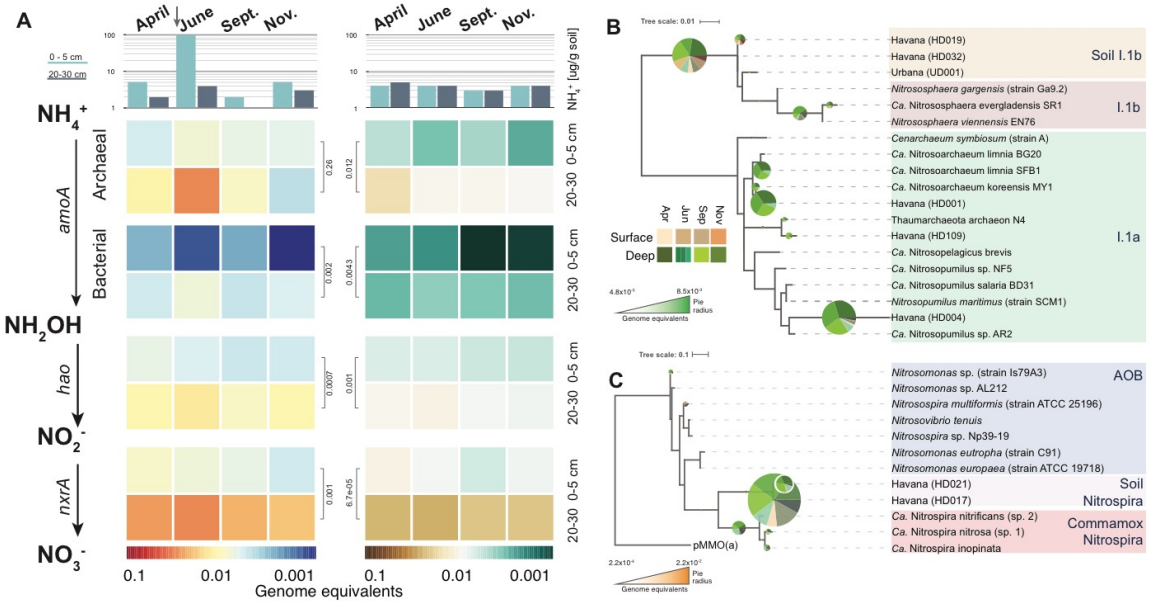


Figure 4.2. Abundance and diversity of nitrification genes in sandy (Havana) and silt-loam (Urbana) soils.

(A) Top panel shows the concentration of NH_4^+ at both sites and depths. Arrow shows the point in time when nitrogen fertilizer (UAN28, 180 lb N/Acre) was applied in Havana. Heatmaps represent calculated relative abundance of nitrification genes (genome equivalents) for Havana (left panel) and Urbana (right panel) soil samples. Values for the 20-30 cm layer in June represent the average

of the three soil cores. Right panels show the phylogenetic reconstruction of archaeal (**B**) and bacterial (**C**) AmoA protein sequences recovered from contigs. Names in parentheses indicate the corresponding metagenomic bins. Both trees include reference protein sequences and assembled sequences from both soil metagenomes. The pie charts represent the placing of Havana metagenomic reads for archaeal and bacterial *amoA* genes using RAxML EPA. Pie chart radii represent the read abundance for each node (calculated as genome equivalents) and the colors of the slices represent the depth and month for metagenomic reads originated from.

In Havana, 60.5% of archaeal *amoA* reads were placed within the I.1a clade (Figure 4.2b). Notably, 55% of the bacterial *amoA* reads were placed within a clade closely related to the recently described Comammox *Nitrospira* (or soil *Nitrospira* for simplicity; Figure 4.2b). Only 15% of the total bacterial *amoA* reads where placed within the *Betaproteobacteria amoA*, and the latter sequences were mostly derived from the top layer. As expected, similar results were observed for *hao* and most *nxrA* gene fragments belonged to different NOB clades (Appendix C, Figure C.6). On the other hand, in Urbana, the majority of the archaeal *amoA* gene fragments (87.3%) were placed within the I.1b clade whereas less than 10% were placed in the I.1a clade (Appendix C, Figure C.7). Compared to Havana, fewer bacterial *amoA* gene fragments were detected in Urbana (Figure 4.2). A similar fraction (58.9%) of the *amoA* gene fragments were placed within the soil *Nitrospira* clade but a higher fraction (~39.7%) was placed inside *Betaproteobacteria* clades. Similar to Havana, the majority of the *hao* and *nxrA* gene fragments were placed in soil *Nitrospira* clades (Appendix C, Figure C.7).

4.3.4 Spatiotemporal Abundance of Population Bin Genomes

To precisely identify and quantify individual populations and their dynamics throughout the season, we performed genome binning analysis of the metagenomic datasets. Additional details about recovered bins can be found in the supporting information section. The abundance of each recovered genome bin, calculated as the fraction of total short-reads mapping on all bins, was stable over time for most of bins from both field sites (Appendix C, Figure C.8). In fact,

only from ~9% to 24% of the total bins (n=69) showed 2-fold change in abundance between any two sampling points. For instance, from April to June, bins HD109 (nitrifier *Thaumarchaeota*), HD116 (*Bacteroidetes*), and HD098 (*Acidobacteria*) showed over 2-fold increase and only bin UD001 showed a similar amount of decrease in abundance. Interestingly, *Thaumarchaeota* nitrifier bins HD109, HD001, and HD004 in addition to bins HD098, HD116 and HD051 showed more than 2-fold decrease from June to September. Lastly, from September to November while only two bins (HD051 and HD103) showed a slight increase over 2-fold, 14 bins showed more than 2-fold decrease, three of which were *Thaumarchaeota* bins (HD032, HD019, and HD109). Other comparisons between April and September (i.e., no crops vs. mature crops) or April and November (i.e., no crops), indicated that 22% and 26% of bins experienced above 2-fold changes, respectively. In these two comparisons *Thaumarchaeota* bins HD001 and HD004 showed more than 3-fold increase indicating the prolonged effect of fertilization over these archaeal populations and/or the high persistence of these organisms during changing environmental conditions.

4.3.5 Diversity of Population Bin Genomes Involved in Nitrogen Cycling

The examination of key N-cycling genes showed that 39 bins, representing different bacterial and archaeal lineages, encoded hallmark nitrification and denitrification enzymes (Figures 4.3a and 4.3b). The latter bins, mostly recovered from Havana, represented almost 40% of the bins showing 2-fold increase or decrease in abundance at any sampling point (Appendix C,

Figure C.8). Bins encoding nitrification enzymes mostly belonged to *Thaumarchaeota* and *Nitrospirae* phyla. All *Thaumarchaeota* bins had at least one copy of the *amoA* and *ureC* genes. Interestingly, all bacterial nitrifier bins also contained the *ureC* gene in addition to all the necessary genes associated with complete oxidation of ammonium to nitrate, i.e., *amoA*, *hao*, *nxrA* genes, similarly to the recently described *Nitrospira* organisms capable of performing complete nitrification (9, 10). On the other hand, none of the bins encoded all the genes to perform canonical or complete denitrification (i.e., reduction of NO_3^- or NO_2^- to N_2), indicating that complete denitrifiers are not abundant in these soils. Instead, most bins obtained from both sites encoded single steps of the denitrification pathway (i.e., Figures 4.3a and 4.3b). Notably, *Nitrospirae* and *Thaumarchaeota* bins, commonly associated with ammonia-oxidizing activity, showed increased abundance upon N-fertilization in Havana (Figure 4.3c), suggesting their participation in nitrification in the agricultural soils.

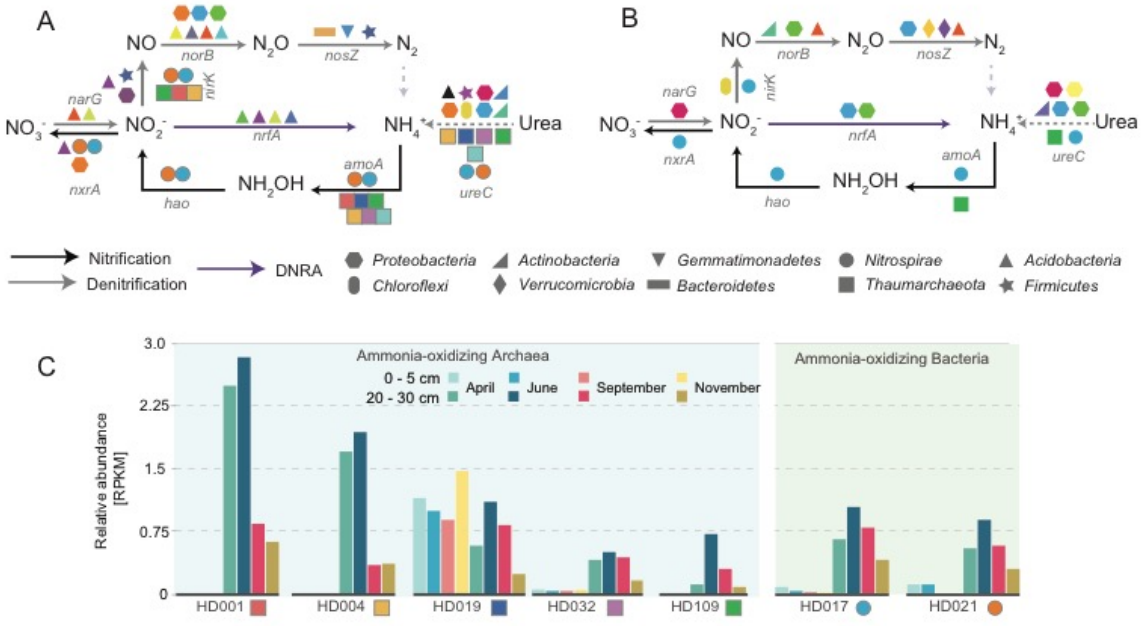


Figure 4.3. Nitrogen cycle genes present in selected population bins and population abundance dynamics across the year in Havana.

(**A** and **B**) Metagenomic bins obtained from Havana (A) and Urbana (B) were queried for the presence of hallmark nitrogen cycle gene markers using HMM models. Each color represents an individual metagenomic bin and the shape depicts the predicted taxonomy at phylum level. Arrows show predicted nitrogen cycle pathways. (C) Bar plots show the relative abundance (y-axis, measured as reads per kilobase per million reads, or RPKM) for nitrifier bins (x-axis) obtained from Havana soil metagenomes. Bright colors represent samples from the 0-5 cm soil layer whereas darker color equivalent corresponds to samples from the 20-30 cm layer (see figure key).

Whole-genome phylogenetic analyses showed that these probable nitrifier bins represented novel taxa (Figure 4.4). For instance, ten bins encoding divergent archaeal and bacterial *amoA* genes were recovered from both sites. Based on average amino acid (AAI) comparisons (14), these ammonia-oxidizing bins most likely represent new genera within the *Thaumarchaeota* and *Nitrospirae* phyla (Appendix C, Table C.6). Consequently, phylogenetic reconstruction based on concatenated single-copy proteins revealed that archaeal bins HD032, HD019, and HT001 from Havana and UD001 from Urbana formed an independent sister clade of the *Thaumarchaeota* group I.1b, mostly consisting of *Thaumarchaeota* isolated from soil (Figure 4.4a). The remaining archaeal bins were placed within the 1.1a group, although bin HD004 formed a deep branch within this group. Other ammonia-oxidizing archaeal bins from Havana such as HD001 were placed close to *Nitrosoarchaeum limnia* strains whereas HD109 was clustered with *Nitrosotenuis cloacae*. Bacterial bins from Havana HD017 and HD021 and Urbana UD063, formed an independent clade closely related to the recently described Comammox *Candidatus* Nitrospira inopinata. In fact, AAI values between *Ca.* N. inopinata and bins HD017, HD021 and UD063 were 65.28%, 66.26% and 65.78%, respectively, indicating their similarity at the genus level. Even though the Havana bin HD021 and the Urbana bin UD063 were obtained from different agricultural sites, they shared 96.09% average nucleotide identity (ANI, SD: 3.49%, from 5,838/12,667 1 Kb-long fragments), revealing that they represent closely related populations, at the level of (same) species (15). Independent phylogenetic trees using hallmark

nitrification proteins AmoA, HAO and NxrA (Appendix C, Figures C.6 and C.7) showed a similar topology to the one observed when using concatenated alignment of multiple single-copy proteins (Figure 4.4), indicating limited recent horizontal gene transfer of the genes. Interestingly, the genetic context of the *amoCAB* operon differed in all bins (Figure 4.4c). For instance, even though HD017 did not encode a copy of the *amoC* gene in the same contig (probably missed during assembling), the synteny of *amoA*, *amoB*, the genes *copD* and *copC* encoding copper binding proteins, and *amoD2*, *amoD1* and *amoE1* showed the same arrangement found in *Ca.* N. inopinata. On the other hand, the *copD* and *copC* genes were absent in bins UD063 and HD021. Intriguingly, bin HD021 possessed a copy of the nitrite reductase (*nirK*) gene, a hallmark gene of denitrification, upstream the *amoCAB* operon, which was surrounded by transposase and integration elements.

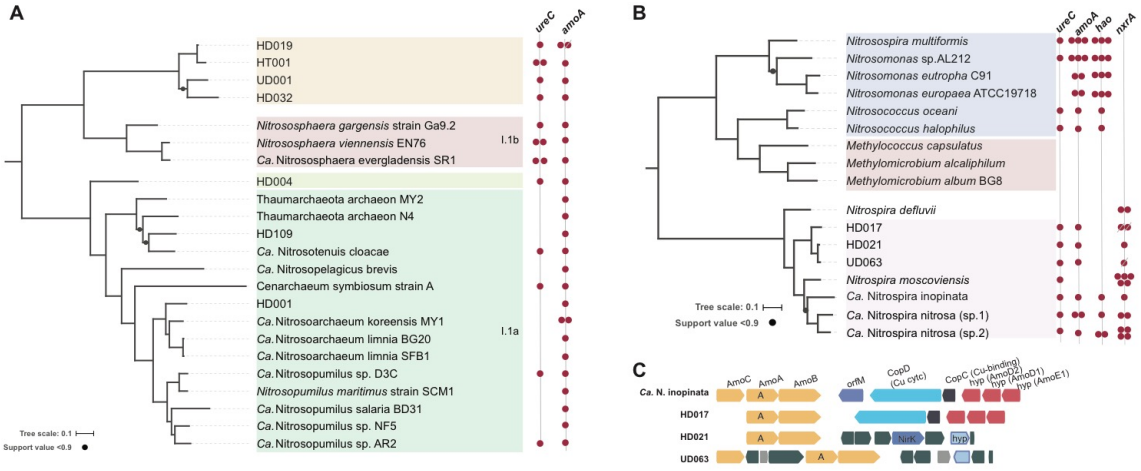


Figure 4.4. Recovery of indigenous archaeal and bacterial ammonia-oxidizing populations.

Phylogenetic reconstruction of archaeal (**A**) and bacterial (**B**) genomes bins encoding *amoA* genes. Concatenated alignments of conserved genes for bacterial or archaeal genomes were used to build maximum likelihood trees in RAxML. Colored circles on the right of each tree show the presence of nitrification gene markers. Strikethrough circles indicate incomplete sequences detected in metagenomic bins. Panel **C** shows a comparison of the genetic context for the *amoA* genes found in *Ca.* N. inopinata and nitrifier bacterial bins recovered from Havana and Urbana.

4.4 DISCUSSION

4.4.1 Temporal Stability of Natural Microbial Communities During the Growing Season

The soil metagenomes obtained at two different depths and four time points throughout the year, from two agricultural fields, which received different management, offered a unique opportunity to explore the functional and community dynamics of indigenous microbial communities. Our results revealed a remarkable composition stability for these microbial communities in their functional, taxonomic, and individual population components during the sampled period. Consistent with the findings reported here, previous reports focusing on single phylogenetic markers (e.g., 16S rRNA gene amplicons) have identified stable genetic composition across soils having different land use or agricultural practices (16, 17), but lacked resolution at the functional gene and individual population levels. It is important to note that the metagenomic snapshots reported here might have missed short-term abundance dynamics or gene expression-activity shifts. Nonetheless, it is clear that the seasonal shifts of soil microbial communities observed were much less profound compared to freshwater or ocean ecosystems, even when the differences in sampling procedures and sample heterogeneity were taken into account. For instance, freshwater metagenomes obtained over 1 year from Lake Lanier showed ~7.5 - fold higher coefficients of variation in gene functions or taxonomic composition, on average, compared to the soil metagenomes (Appendix C, Figure C.2c), highlighting that seasonality has a stronger effect in structuring the lake microbial

communities (18, 19). Further, larger taxa and gene-content differences were observed, in general, between spatial (e.g., different depths) relative to temporal (e.g., four seasons) scales for both agricultural soils. These findings suggest that the biotic (e.g., plants) and edaphic soil physicochemical characteristics have a stronger effect in structuring and modulating the variability of microbial soil communities compared to seasonal changes, in agreement with previous findings (20). These findings suggest that many soil organisms may withstand changing conditions by modulating specific gene expression rather than undergoing changes in their abundance and/or are well-adapted to the seasonal environmental fluctuations in soils (e.g., carbon and nitrogen inputs). In fact, soil bacteria have larger genome sizes compared to other environments (21) and harbor a greater variety of metabolic pathways compared to their water or human-associated counterparts (22, 23), which is consistent with the latter interpretations.

4.4.2 Impact of N-fertilizer on Microbial Soil Communities

While most populations examined here showed steady abundance throughout the seasons, a fraction showed conspicuous responses to agricultural activities. The latter populations showed 2-fold change, or higher, which represents a substantial change in abundance for the slow growing conditions that prevail in bulk soil (e.g., 1-2 generations per year, on average) (24). The largest changes in abundance observed at any time point sampled were clearly those in response to synthetic nitrogen fertilization in the Havana site. Previously unrecognized nitrifier populations belonging to the *Thaumarchaeota* phylum,

showed an increase and subsequent decrease in abundance upon the application of N fertilizers. The latter microbial populations might also potentially act as a source of N_2O as previously suggested (7, 25). In Urbana, probably because this site did not receive synthetic fertilizer in the sampling year, many bins showed stable abundances, meaning that they did not exhibit changes in abundance greater than 2-fold. In addition, the high abundance of bins encoding enzymes necessary for oxidation (i.e., nitrification) or reduction (i.e., denitrification) of N species underscored the effect of N-fertilization on a selected fraction of the microbial communities involved in N-cycle. These findings also indicated that most community members may be well adapted to the seasonal and soil depth-specific environmental changes in temperature and other variables and thus, did not change much in abundance relative to the nitrifiers.

Interestingly, none of the recovered bins encoded all enzymes required for performing complete denitrification. Even though the 69 recovered bins are far from representing extant soil microbial community diversity, they most likely better represent the most abundant organisms at the times samples, whose genome was recoverable by assembly and binning. Previous reports focusing on organisms harboring clade II or atypical *nosZ* genes obtained from the same agricultural sites have also proposed a modular assembly for denitrification pathways in these soils (26, 27), consistent with the results reported here. These findings suggest that reduction of oxidized N species to nitrogen gas (N_2) would require the concerted participation of different N-reducing organisms, and highlights the importance of accounting for the latter organisms and their

interaction in better understanding denitrification processes and the nitrogen balance (e.g., N-losses in the form of N_2O gas) in soil ecosystems.

4.4.3 Novel Nitrifiers in Agricultural Soils

Even though previous reports have detected high abundance of common nitrification marker genes or 16S rRNA gene sequences assignable to known nitrification taxa in soils (11, 28), the genomic information of these taxa has been limited. Both agricultural sites examined here showed high abundances of *Nitrospirae* and *Thaumarchaeota* communities despite their different edaphic characteristics (but similar crop rotation management). The bacterial nitrifier genomes recovered from the two sites represented divergent clades from those of the well-characterized and canonical *Betaproteobacteria* and *Gammaprotebacteria* nitrifiers. In fact, these soil organisms were most closely related, yet distinct at the genus level based on their AAI relatedness, to the novel Comammox *Nitrospira*, which was recently demonstrated being capable of performing complete nitrification (9, 10). The latter organisms were isolated from a pipe and trickling filter biofilms, with contrasting environmental conditions and N inputs compared to agricultural soils. Therefore, it appears that the agricultural management in the two soils has selected for discrete-evolving nitrifiers population with different ecophysiologies in addition to the known nitrifiers. In addition, the substantially different physicochemical characteristics of the two soils examined indicated that the novel *Nitrospira* and *Thaumarchaeota* recovered genomes may represent ecologically successful populations, at least for agricultural soils, and could serve as models for future studies.

Previous reports have proposed an oligotrophic nature for AOA based on AmoA substrate affinity (29), and suggested that they are strongly enriched in specific soil horizons (30). Our study showed that these archaeal organisms might also thrive in environments receiving large inputs of N and, perhaps more importantly, they respond to N fertilization and do not follow the predicted oligotrophic lifestyle. In fact, archaeal genomic bins change in abundance as much as their bacterial counterparts, which are thought to be slower growers (K-strategists). These findings indicted that, in soils receiving yearly large inputs of synthetic N fertilization, indigenous microbes, including AOA, have been adapted to these conditions and thus, evolved to be more r-strategists than their counterparts in non-agricultural soils. Nonetheless, this hypothesis awaits experimental verification by examining activity rates and substrate affinities of indigenous nitrifier bacteria and archaea.

The five abundant archaeal nitrifier genomes recovered also showed distinct distributions with depth. For instance, bin HD109, was the only AOA population showing elevated abundance in top and deeper soil layers, whereas the rest of *Thaumarchaeota* bins showed high abundance only in deep soil layers. Apparently, additional niches and ecophysiologies that remain to be elucidated underlay the distribution patterns for the detected AOA. On the other hand, Comammox nitrifier populations represented by bins from Havana (HD021) and Urbana (UD063) showed high level of similarity (ANI >95%), revealing ecological success in agricultural soils with contrasting characteristics. These nitrifier genes and genomes were likely missed by previous studies that

employed probes designed based on available sequences of functional (e.g., AmoA) or 16S rRNA gene sequences (31). Collectively, our results propose a role for novel AOA and Comammox organisms in responding and oxidizing high N inputs from fertilization and provided the means, e.g., genome sequences, to track the abundant nitrifier populations in agricultural soils in Midwest US.

Altogether, our study showed stable microbial communities dwelling in agricultural soils and identified key populations and genes responding to seasonal (e.g., fall biomass return) and human-induced perturbations (e.g., fertilization practices). These findings also propose a much broader niche for the recently described Comammox organisms and ammonia-oxidizing archaea controlling the fate of nitrogen in agricultural soils.

4.5 EXPERIMENTAL PROCEDURES

4.5.1 Soil Samples and DNA Extraction and Sequencing

Agricultural soil samples were collected in 2012 from Havana, IL (lat 40.296, long 89.944; elevation, 150 m) and Urbana, IL (lat 40.075, long 88.242; elevation, 222 m) field locations both with long histories of conventionally managed corn and soybean rotations. The Havana field site is characterized as a sand (93% sand; 7% clay) with somewhat excessive drainage with no ponding duration or frequency. During the summer season, the field is irrigated with underlying groundwater from the Mahomet aquifer. The Urbana field site is situated on a slight slope profile, characterized as a silt loam (average content 21% sand; 69% silt; 10% clay). This site is classified as poorly drained, with brief

and frequent ponding, and soil moisture is exclusively due to precipitation events. Once each season (April, June, Late August/Early September and November), three soil cores (2.5 cm width by 30 cm length) were collected at three fixed locations 30 m apart (centroids) within each field plot (9 cores total per field, per sampling time), with each core then partitioned into two depths (0-5 cm and 20-30 cm). In sampling year (2012), corn was seeded in Havana and received UAN28 (28%N as urea:ammonium:nitrate) fertilizer (180 lb N/Acre), whereas Urbana was planted with soybean and no fertilizer was applied (additional soil management events conducted in 2012 are described in Appendix C, Table C.2). Soil physicochemical characteristics (organic matter, P, K, Mg, Ca, NO_3^--N, NH_4^+-N, total N, CEC) were determined using a composite pool of soil combined for each depth range at each time of sampling (A&L Laboratories, Ft. Wayne, IN) (Appendix C, Table C.1). Soil pH and gravimetric soil moisture content were measured using independent fractions from each soil core. DNA was extracted from ~0.5 g of soil from each fraction using a modified phenol-chloroform and purification protocol as previously described (27), and equal quantities of DNA from each fraction based on agarose gel quantification were pooled to create composite samples for each of the two soil depths (0-5 cm and 20-30 cm) for each soil type. The 20-30 cm soil sections taken from both sites in June were pooled according to centroid (Havana- E, M, and W; Urbana-N, M, and S) and these pools were independently sequenced. Sequencing of DNA samples were performed using the Illumina HiSeq 2000 platform and the Nextera DNA 150x150 library preparation protocol (Appendix C, Table C.3) as described previously (27).

4.5.2 Short-Read Assembly and Analyses

Metagenomic raw reads (FASTQ) for all samples were trimmed using SolexaQA (32) using a Phred score cutoff of 20 and minimum fragment length of 50 bp. Average coverage for each sequenced library was determined by Nonpareil (13) using default settings (Appendix C, Table C.3). Protein-coding sequences were predicted from the short-read metagenomes using FragGeneScan (33) and functional annotation was performed by BLASTp v2.2.29+ searches (34) against UniProt (35) using default parameters. BLAST search outputs were filtered for best match and minimum identity >=50% and read alignment >=70%. Protein annotations were subsequently translated into SEED subsystems (36) for functional analyses. Calculated MASH distances (37) and annotation counts from SEED subsystems were visualized in ordination plots (NMDS) using the vegan (38) and ecodist (39) libraries in R v3.3.1. Differentially abundant SEED functional annotations or taxonomical levels were determined with the DESeq2 (40) package. The homogeneity of the variance (i.e., homoscedasticity) across groups was corroborated by using Levene's test implemented in the car package (41) and Bartlett's test available in R. Short-read metagenomes were co-assembled as Havana top (four samples), Havana deep (six samples), Urbana top (four samples) and Urbana deep (six samples) using IDBA_UD v1.1.1 (42) (Appendix C, Table C.4).

4.5.3 Identification of Nitrogen Cycle Genes

To identify and quantify reads encoding specific proteins of interest, in-house databases were constructed and manually curated using sequences obtained from UniProt (35) for the archaeal and bacterial ammonia monooxygenase alpha subunit (AmoA), hydroxylamine oxidase (Hao), nitrite oxidoreductase alpha subunit (NxrA), nitrate reductase (NarG), nitrite reductase (NirK), nitric oxide reductase beta subunit (NorB), nitrous oxide reductase (NosZ), nitrite reductase (NrfA) and DNA-directed RNA polymerase subunit beta (RpoB) (Appendix C, Table C.5). Independent ROCker (43) models (length=125 bp, as it was the average for all the metagenomes) were subsequently built based on these databases with the exception of NarG and NxrA, where the databases were combined as a single model (Appendix C, Table C.5). Trimmed short-reads from soil metagenomes were used as query for BLASTx searches (evalue 0.01) against the latter protein databases and outputs were filtered using the previously generated ROCker models. Target gene abundance was determined as genome equivalents by calculating the ratio between normalized target reads (counts divided by median protein length) and normalized RpoB reads (counts divided by median RpoB protein length). Protein databases and ROCker models are available through http://enve-omics.ce.gatech.edu. We also searched for assembled N-cycle protein sequences in the four co-assemblies and bins using precompiled hidden Markov models obtained from FUNGENE (44) and HMMer (45). Detected target N-cycle proteins were manually curated by assessing the presence of characteristic amino acid and phylogenetic congruency. Recovered AmoA sequences were also used for constructing

phylogenetic trees. Construction of phylogenetic trees and short-read placement and visualization were performed as previously reported (43) and additional details are available in supporting information.

4.5.4 Recovery of Metagenomic Bins and Analyses

Assembled contigs over 1,000 bp were binned from each co-assembly using an expectation-maximization algorithm as implemented in MaxBin v2.1.1 (46). Bins over 70% completion according to CheckM (47) were re-assembled using the reads mapping on the contigs of the bin (identity => 98% and fraction of aligned read >=70%) in Velvet v1.2.10 (48). Taxonomic classification and novelty for the obtained bins were assessed in Microbial Genome Atlas (MiGA) (http://enve-omics.ce.gatech.edu:3000). Contigs with conflicting taxonomy (most likely representing binning misplacement) were manually inspected by comparing calculated coverage and inferred taxonomical classifications from MyTaxa (49) scan reports as implemented in MiGA and subsequently manually discarded (Appendix C, Table C.6). Reported bin statistics (contamination and completeness) were determined using the “HMM.essential.rb” script from the enveomics collection and manually corrected for fragmented single-copy genes (e.g., a single-copy gene fragmented into two different contigs). Bin abundance was calculated as reads per kilobase per million reads (RPKM) using the matching reads to the binned contigs from BLASTn searches (identity >=98% and fraction of read aligned >= 50%) divided by the metagenomic sample sizes (in millions of reads) and the length of the bin genome in kilo bases. Phylogenetic reconstruction of bins was performed based on concatenated alignments of

universal single-copy proteins identified for each bin using the "HMM.essential.rb" script. Forty bacterial and nine archaeal proteins present in the corresponding bins were extracted and multiple alignments for each protein were generated using ClustalΩ. Concatenated alignments without invariable sites were generated for archaeal and bacterial alignments using the script "Aln.cat.rb" from the enveomics collection. Phylogenetic reconstructions were determined using in RAxML v8.0.19 (-f a, -m PROTGAMMAAUTO, –N 100) and visualized in iTol.

4.5.5 Data Availability

Raw metagenomic soil samples are deposited in the European Nucleotide archive under study number PRJEB20068.

4.6 REFERENCES

1. **Ravishankara AR**, **Daniel JS**, **Portmann RW**. 2009. Nitrous oxide (N2O): the dominant ozone-depleting substance emitted in the 21st century. Science **326**:123–125.

2. **Seitzinger SP**, **Kroeze C**, **Styles RV**. 2000. Global distribution of N2O emissions from aquatic systems: natural emissions and anthropogenic effects. Chemosphere - Global Change Science **2**:267–279.

3. **Firestone MK**, **Davidson EA**. 1989. Microbiological basis of NO and N_2O production and consumption in soil, pp. 7–21. *In* Andreae, MO, Schimel, DS, Robertson, GP (eds.), Exchange of trace gases between terrestrial ecosystems and the atmosphere. Wiley and Sons: New York.

4. **Bremner JM**. 1997. Sources of nitrous oxide in soils. Nutrient Cycling in Agroecosystems **49**:7–16.

5. **Luo C**, **Rodriguez-R LM**, **Johnston ER**, **Wu L**, **Cheng L**, **Xue K**, **Tu Q**, **Deng Y**, **He Z**, **Shi JZ**, **Yuan MM**, **Sherry RA**, **Li D**, **Luo Y**, **Schuur EAG**, **Chain P**, **Tiedje JM**, **Zhou J**, **Konstantinidis KT**. 2014. Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. Appl Environ Microbiol **80**:1777–1786.

6. **Hug LA**, **Thomas BC**, **Sharon I**, **Brown CT**, **Sharma R**, **Hettich RL**, **Wilkins MJ**, **Williams KH**, **Singh A**, **Banfield JF**. 2015. Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. Environmental Microbiology **18**:159–173.

7. **Prosser JI**, **Nicol GW**. 2012. Archaeal and bacterial ammonia-oxidisers in soil: the quest for niche specialisation and differentiation. Trends in Microbiology **20**:523–531.

8. **Jung M-Y**, **Well R**, **Min D**, **Giesemann A**, **Park S-J**, **Kim J-G**, **Kim S-J**, **Rhee S-K**. 2013. Isotopic signatures of N2O produced by ammonia-oxidizing archaea from soils. The ISME Journal **8**:1115–1125.

9. **Daims H**, **Lebedeva EV**, **Pjevac P**, **Han P**, **Herbold C**, **Albertsen M**, **Jehmlich N**, **Palatinszky M**, **Vierheilig J**, **Bulaev A**, **Kirkegaard RH**, **Bergen von M**, **Rattei T**, **Bendinger B**, **Nielsen PH**, **Wagner M**. 2015. Complete nitrification by Nitrospira bacteria. Nature **528**:504–509.

10. **van Kessel MAHJ**, **Speth DR**, **Albertsen M**, **Nielsen PH**, **Op den Camp HJM**, **Kartal B**, **Jetten MSM**, **Lücker S**. 2015. Complete nitrification by a single microorganism. Nature **528**:555–559.

11. **Verhamme DT**, **Prosser JI**, **Nicol GW**. 2011. Ammonia concentration determines differential growth of ammonia-oxidising archaea and bacteria in soil microcosms. The ISME Journal **5**:1067–1071.

12. **Levičnik-Höfferle Š**, **Nicol GW**, **Ausec L**, **Mandić-Mulec I**, **Prosser JI**. 2012. Stimulation of thaumarchaeal ammonia oxidation by ammonia derived from organic nitrogen but not added inorganic nitrogen **80**:114–123.

13. **Rodriguez-R LM**, **Konstantinidis KT**. 2014. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics **30**:629–635.

14. **Konstantinidis KT**, **Tiedje JM**. 2007. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Current Opinion in Microbiology **10**:504–509.

15. **Goris J**, **Konstantinidis KT**, **Klappenbach JA**, **Coenye T**, **Vandamme P**, **Tiedje JM**. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol **57**:81–91.

16. **Lauber CL**, **Ramirez KS**, **Aanderud Z**, **Lennon J**, **Fierer N**. 2013. Temporal variability in soil microbial communities across land-use types. The ISME Journal **7**:1641–1650.

17. **Hartmann M**, **Frey B**, **Mayer J**, **Mäder P**, **Widmer F**. 2015. Distinct soil microbial diversity under long-term organic and conventional farming. The ISME Journal **9**:1177–1194.

18. **Shade A**, **Kent AD**, **Jones SE**, **Newton RJ**, **Triplett EW**, **McMahon KD**. 2007. Interannual dynamics and phenology of bacterial communities in a eutrophic lake. Limnology and Oceanography **52**:487–494.

19. **Crump BC**, **Peterson BJ**, **Raymond PA**, **Amon RMW**, **Rinehart A**, **McClelland JW**, **Holmes RM**. 2009. Circumpolar synchrony in big river bacterioplankton. Proc Natl Acad Sci USA **106**:21208–21212.

20. **Lauber CL**, **Hamady M**, **Knight R**, **Fierer N**. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol **75**:5111–5120.

21. **Konstantinidis KT**, **Tiedje JM**. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci USA **101**:3160–3165.

22. **Bentley SD**, **Parkhill J**. 2004. Comparative genomic structure of prokaryotes. Annu Rev Genet **38**:771–792.

23. **Ley RE**, **Peterson DA**, **Gordon JI**. 2006. Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine. Cell **124**:837–848.

24. **Gray T**, **Williams ST**. 1971. Microbial productivity in soil. Cambridge University Press, Cambridge.

25. **Stieglmeier M**, **Alves RJE**, **Schleper C**. 2014. The Phylum Thaumarchaeota, pp. 347–362. *In* Rosenberg, E, DeLong, EF, Lory, S, Stackebrandt, E, Thompson, F (eds.), The Prokaryotes. Springer Berlin Heidelberg, Berlin, Heidelberg.

26. **Sanford RA**, **Wagner DD**, **Wu Q**, **Chee-Sanford JC**, **Thomas SH**, **Cruz-García C**, **Rodríguez G**, **Massol-Deyá A**, **Krishnani KK**, **Ritalahti KM**, **Nissen S**, **Konstantinidis KT**, **Löffler FE**. 2012. Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils. **109**:19709–19714.

27. **Orellana LH**, **Rodriguez-R LM**, **Higgins S**, **Chee-Sanford JC**, **Sanford RA**, **Ritalahti KM**, **Löffler FE**, **Konstantinidis KT**. 2014. Detecting nitrous oxide reductase (NosZ) genes in soil metagenomes: method development and implications for the nitrogen cycle. mBio **5**:e01193–14.

28. **Leininger S**, **Urich T**, **Schloter M**, **Schwark L**, **Qi J**, **Nicol GW**, **Prosser JI**, **Schuster SC**, **Schleper C**. 2006. Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature **442**:806–809.

29. **Martens-Habbena W**, **Berube PM**, **Urakawa H**, **la Torre de JR**, **Stahl DA**. 2009. Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. Nature **461**:976–979.

30. **Jung M-Y**, **Park S-J**, **Kim S-J**, **Kim J-G**, **Damsté JSS**, **Jeon CO**, **Rhee S-K**. 2014. A Mesophilic, Autotrophic, Ammonia-Oxidizing Archaeon of Thaumarchaeal Group I.1a Cultivated from a Deep Oligotrophic Soil Horizon. Appl Environ Microbiol **80**:3645–3655.

31. **Santoro AE**. 2016. The do-it-all nitrifier. Science **351**:342–343.

32. **Cox MP**, **Peterson DA**, **Biggs PJ**. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics **11**:485.

33. **Rho M**, **Tang H**, **Ye Y**. 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res **38**:e191.

34. **Camacho C**, **Coulouris G**, **Avagyan V**, **Ma N**, **Papadopoulos J**, **Bealer K**, **Madden TL**. 2009. BLAST+: architecture and applications. BMC Bioinformatics. BioMed Central Ltd.

35. **UniProt Consortium**. 2015. UniProt: a hub for protein information. Nucleic Acids Res **43**:D204–12.

36. **Overbeek R**, **Begley T**, **Butler RM**, **Choudhuri JV**, **Chuang H-Y**, **Cohoon M**, **de Crécy-Lagard V**, **Diaz N**, **Disz T**, **Edwards R**, **Fonstein M**, **Frank ED**, **Gerdes S**, **Glass EM**, **Goesmann A**, **Hanson A**, **Iwata-Reuyl D**, **Jensen R**, **Jamshidi N**, **Krause L**, **Kubal M**, **Larsen N**, **Linke B**, **McHardy AC**, **Meyer F**, **Neuweger H**, **Olsen G**, **Olson R**, **Osterman A**, **Portnoy V**, **Pusch GD**, **Rodionov DA**, **Rückert C**, **Steiner J**, **Stevens R**, **Thiele I**, **Vassieva O**, **Ye Y**, **Zagnitko O**, **Vonstein V**. 2005. The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. Nucleic Acids Res **33**:5691–5702.

37. **Ondov BD**, **Treangen TJ**, **Melsted P**, **Mallonee AB**, **Bergman NH**, **Koren S**, **Phillippy AM**. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol **17**:403.

38. **Oksanen J**, **Kindt R**, **Legendre P**, **O'Hara B**. 2007. vegan: Community ecology package. Community ecology package **10**:631–637.

39. **Goslee SC**, **Urban DL**. 2007. The ecodist package for dissimilarity-based analysis of ecological data. Journal of Statistical Software **22**:1–19.

40. **Love MI**, **Huber W**, **Anders S**. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **15**:550.

41. **Fox J**, **Weisberg S**. 2011. An R Companion to Applied Regression, 2nd ed. SAGE Publications, Los Angeles, USA.

42. **Peng Y**, **Leung HCM**, **Yiu SM**, **Chin FYL**. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28**:1420–1428.

43. **Orellana LH**, **Rodriguez-R LM**, **Konstantinidis KT**. 2017. ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. Nucleic Acids Res **45**:e14.

44. **Fish JA**, **Chai B**, **Wang Q**, **Sun Y**, **Brown CT**, **Tiedje JM**, **Cole JR**. 2013. FunGene: the functional gene pipeline and repository. Frontiers in Microbiology **4**:291.

45. **Eddy SR**. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol **7**:e1002195.

46. **Wu Y-W**, **Tang Y-H**, **Tringe SG**, **Simmons BA**, **Singer SW**. 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome

2015 3:1 **2**:26.

47. **Parks DH**, **Imelfort M**, **Skennerton CT**, **Hugenholtz P**, **Tyson GW**. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res **25**:1043–1055.

48. **Zerbino DR**, **Birney E**. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18**:821–829.

49. **Luo C**, **Rodriguez-R LM**, **Konstantinidis KT**. 2014. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. Nucleic Acids Res **42**:e73–e73.

CHAPTER 5. COMPARISON OF MULTI-OMICS FOR PREDICTING THE RATE OF MICROBIAL NITROGEN UTILIZATION IN SOILS

Reproduced in part with permission from Luis H. Orellana, Janet K. Hatt, Karuna Chourey, Robert L. Hettich, Wendy Yang, Joanne C. Chee-Sanford, Robert A. Sanford, Frank E. Löffler, and Konstantinos T. Konstantinidis. Comparison of multi-omics for predicting the rate of microbial nitrogen utilization in soils. All copyright interests will be exclusively transferred to the publisher upon submission.

5.1 ABSTRACT

Multi-omic techniques can offer a comprehensive overview of microbial communities at the gene, transcript and protein levels. However, to what extent these approaches can predict *in-situ* microbial activity is less clear, especially in highly complex habitats such as soil. Here we performed microcosm incubations using soil from a site with a long history of agricultural management. The microcosms were amended with ammonium and urea, simulating a fertilization event, and showed high nitrification rate (~2 μNO_3^--N g^{-1} soil d^{-1}) after 2 days and accumulation of N_2O after 8 days. Nitrification activity ($NH_4^+ \rightarrow NH_2OH \rightarrow NO_2^- \rightarrow NO_3^-$) was accompanied by 6-fold or more increase in relative expression ratios (cDNA/DNA) for *Nitrosomonas* and *Nitrospira* between 10h and 8 days of incubation. In contrast, archaeal *Nitrososphaera* and *Nitrosopumilus*, and Comammox *Nitrospira* nitrifiers showed stable expression during the incubations although they were generally more abundant (DNA level) than their betaproteobacterial counterparts. A strong linear relationship (R^2>0.95) was

observed between nitrification activity and ammonia monooxygenase (*amoA*; $NH_4^+ \rightarrow NH_2OH$) and nitrite oxidoreductase (*nxrA*; $NO_2^- \rightarrow NO_3^-$) gene or transcript abundances in time-series samples, revealing that both DNA and mRNA levels quantitatively reflected activity. Although peptides related to housekeeping and NrxA proteins from nitrite-oxidizing organisms were detected their abundance was not significantly correlated with activity, revealing that meta-proteomics provided only a qualitative assessment of activity. Altogether, these findings underscored the strengths and limitations of multi-omic approaches for assessing complex microbial communities and provided means to measure nitrification processes in soils.

5.2 INTRODUCTION

Even though the central role of microbes in the cycling of nitrogen (N) is recognized, the dynamics and controls of the interrelated microbial N pathways in agricultural soils are still poorly understood. This scarcity of information limits the development of more accurate, predictive models of N-flux that will encompass the role of microbes in the generation and consumption of N-substrates, as well as the emission of greenhouse gases, including nitrous oxide (N_2O), a potent greenhouse gas (1). In agricultural soils receiving large inputs of N fertilizer, ammonia-oxidizing bacteria (AOB), ammonia-oxidizing archaea (AOA) and nitrite-oxidizing bacteria (NOB) are responsible for the fast conversion of ammonium to nitrate, diminishing the available N to plants. It has also been reported that under low oxygen concentrations, nitrification is a major N_2O source (2), although the microbial enzymes catalyzing these reactions are not fully

understood. Alternatively, under anoxic conditions, nitrate can be reduced to gaseous forms such as N_2, NO or N_2O by denitrifier organisms and consequently leave the soil system. Despite the apparent importance of nitrification in the generation of N_2O, the relative contribution of commamox, AOA, AOB and NOB populations in this process, especially during soil fertilization events, is currently unclear (3). Advancing this issue is essential for better predicting the contribution of these microbial populations to the N-cycle and modeling. High-throughput genomic and proteomic approaches offer the means to characterize the N-pathways in the environment. However, to what extent these omic approaches reflect process rates is still unclear.

Even though DNA, RNA and protein abundances can all reflect microbial activity and responses to environmental changes, each of these levels generally offers different levels of information. For instance, gene sequences (e.g., metagenomics) offer a comprehensive overview of the functional potential of microbial communities but do not generally reflect active community members or functions. Short-term microbial responses to external changes (e.g., N addition) can be best tracked by analyzing the actively expressed genes (i.e., transcript changes). For instance, the relationship between measured nitrification processes and the ammonia monooxygenase (*amoA*) transcripts have elucidated differences between archaeal and bacterial activity in acidic soils (4). Additionally, proteins provide a third level of molecular information much closer to activity by reflecting occurring functional enzymes and reactions. Even though proteomics have been applied to only a limiting number of natural microbial

communities, it has provided new insights about novel metabolic capabilities in highly complex environments (5). Furthermore, recent advances of metagenomics and metaproteomics techniques as well as intergradation with isotope-based technologies (e.g., NanoSIMS) have disentangled the role of previously elusive microbial communities in the environment. For instance, the application of metagenomics and metaproteomics in combination have provided new understanding of novel uncultured microorganisms participating in the cycling of sulfur, nitrogen, and carbon in terrestrial subsurface environments (6). However, only a few reports have examined how these approaches correlate with *in situ* process rates, especially in soil ecosystems that are characterized by high complexity and heterogeneity, and typically low metabolic activity. Metatranscriptomic approaches have been successfully used to examine the degradation of the herbicide atrazine by *Escherichia coli* in bioreactors, showing a linear relationship between the measured enzymatic activity and the transcripts encoding the associated enzyme (7). Additionally, in microbial leaf-litter decomposition incubations, positive correlations were observed between cellulase and xylanase protein abundances and their corresponding enzymatic activities (8). On the other hand, even though the linkage of multi-omic datasets provided new insights into diversity and gene potential of microbial communities of permafrost ecosystems, these datasets were less predictive of measured process rates (9). Therefore, it is not clear to what extent the omic measurements correlate with each other and with *in situ* process rates in soils.

In the present work, we examined N amended sandy soil obtained from a site with long history of agricultural management. Our previous year-round characterization of the same agricultural site revealed responsive novel *Thaumarchaeota* and *Nitrospira* nitrifiers to the field application of synthetic fertilizer, but the findings were limited to metagenomic potential (10). Here, our goal was to assess microbial activity obtained from measurements of DNA, RNA and protein abundances against process rate measurements e.g., nitrate accumulation and N_2O production, in soils incubated under well-control conditions in the laboratory. Our results showed that metatranscriptomic data best reflected the measured nitrification process rates under these conditions, with metagenomics being a close second.

5.3 METHODS

5.3.1 Soil Sampling

Our study was focused on an agricultural plot located in the Havana County, in the State of Illinois, USA (lat 40.296, long 89.944; elevation, 150 m). This site is representative of the US Midwest and has a long history of conventionally managed corn and soybean rotation. In October 2014, we collected ~2kg of bulk soil from the 20-30 cm soil depth as previous results have shown significant presence of ammonia-oxidizing microorganisms in this layer (10). Sample metadata are summarized in Appendix D, Table D.1.

5.3.2 Soil Incubations, Gas and Chemical Analyses

Soil microcosms were established in triplicates, using ~120 gr of soil in 500 ml jars, and were sampled at six time points (0h, 10h, 24h, 48h, 5 days and 8 days). To set up the microcosms, 6 ml of a 40 mM solution of two stable isotopes for NH_4Cl (50% ^{15}N-NH_4Cl and 50% ^{14}N-NH_4Cl) and 20 mM urea (NH_2CONH_2) or 40 mM NH_4Cl and 20 mM of two stable isotopes for urea (50% ^{15}N-NH_2CONH_2 and 50% ^{14}N- NH_2CONH_2) were added to 400 gr of soil. After vigorously mixing, 120 gr of soil were dispensed into three microcosm jars and incubated in a dark growth chamber with diurnal temperature fluctuation as observed in Havana in the spring fertilization period. Un-amended microcosms receiving 6 ml of filtered irrigation water served as controls. After each sampling point, headspace gas was collected and the concentration of O_2, CO_2, N_2O were measured by gas chromatography (GC). Soil was sampled destructively and used for estimating residual ammonium and nitrate concentrations (Mass Spectrometry analysis) using 2M KCl extractions (11). pH and moisture were determined as previously described (Appendix D, Table D.1).$^{15}N_2$,$^{15}N_2O$ and corresponding analytical procedures (GC/MS analyses) were performed as previously described (12). No inhibitors of nitrogen cycle pathways were used in the incubations.

5.3.3 Nucleic Acid Extractions

From each incubation point, DNA was extracted from ~0.5 g of soil using a modified phenol-chloroform and purification protocol as previously described (13). For RNA extraction, 2gr of soil was preserved in LifeGuard (MoBio) and stored at -80C. A modified protocol derived from the PowerMax Soil DNA kit for

extracting RNA was used for total RNA extractions (MoBio). TURBO DNAse (Ambion) was used to remove DNA according to the recommendations of the manufacturer. Nucleic acid extracts were quantified using Quant-it ds DNA HS and HS RNA assays (Invitrogen) according to the instructions of the manufacturer. RNA quality was assessed using Agilent RNA 6000 pico kit (Agilent Technologies) and samples having RNA integrity number (RIN) above 7 were further used.

5.3.4 Nucleic Acid Sequencing

For metagenomes, dual-indexed DNA sequencing libraries were prepared using the Illumina Nextera XT DNA library prep kit according to manufacturer's instructions, except that the protocol was terminated after isolation of cleaned amplified double stranded libraries. For metatranscriptomes, single-indexed cDNA sequencing libraries were prepared using ScriptSeq v2 protocol using ~25 ng of total RNA as input. cDNA library concentrations were determined by fluorescent quantification using a Qubit HS DNA kit and Qubit 2.0 fluorometer (ThermoFisher Scientific) according to manufacturer's instructions and samples were run on a High Sensitivity DNA chip using the Bioanalyzer 2100 instrument (Agilent) to determine quality and average library insert sizes. An equimolar mixture of the libraries was sequenced on an Illumina HiSEQ 2500 instrument (School of Biological Sciences, Georgia Institute of Technology) for a rapid run of 300 cycles (2 x 150 bp paired end) using the HiSeq Rapid PE Cluster Kit v2 and HiSeq Rapid SBS Kit v2 (Illumina). Adapter trimming and

demultiplexing of sequenced samples was carried out by the Illumina software, according to the recommendations of the manufacturer.

5.3.5 Short-read Analyses

Metagenomic and metatranscriptomic raw reads (FASTQ) for all samples were trimmed using SolexaQA (14) using a Phred score cutoff of 20 and minimum fragment length of 50 bp. Short-reads derived from metatranscriptomes were merged using PEAR using default parameters (15). Average coverage for each sequenced metagenome was determined by Nonpareil (16) using default settings except that 2,000 reads were used as query (-X option) (Appendix D, Tables D.2 and D.3).

Short-read sequences encoding 16S rRNA gene fragments were extracted from each metagenome and metatranscriptome by SortMeRNA (17) (Appendix D, Table D.4), and their taxonomy was assigned using RDP classifier (cutoff 50) (18).

To identify and quantify reads encoding specific protein sequences of interest, we used the previously published protein sequences as references (10) for the archaeal ammonia monooxygenase alpha subunit (AmoA), bacterial AmoA, hydroxylamine oxidase (Hao), nitrite oxidoreductase alpha subunit (NxrA), nitrite reductase (NirK), nitric oxide reductase beta subunit (NorB), nitrous oxide (NosZ), nitrite reductase (NrfA) and DNA-directed RNA polymerase subunit beta (RpoB). Independent ROCker (19) models (length=125 bp) were subsequently built based on these reference protein sequences with the exception of NarG and

NxrA, where the sequences were combined into a single model. Trimmed short-reads from soil metagenomes were used as query for BLASTx searches (e-value 0.01) against the latter protein databases and outputs were filtered using the previously generated ROCker models. For metagenomes, target gene abundance in metagenomes was determined as genome equivalents by calculating the ratio between normalized target reads (number of reads matching divided by median protein length) and normalized RpoB reads (number of reads matching divided by median RpoB protein length), a universal single-copy gene. For metatranscriptomes, target transcripts abundance was calculated as reads per kilobase of transcript per million mapped reads (RPKM). Protein databases and ROCker models are available through http://enve-omics.ce.gatech.edu/.

5.3.6 Assembly and Binning of Metagenomic Populations

Short-read metagenomes from all incubation time points were co-assembled using IDBA_UD v1.1.1 and binning was performed as previously described (10). Taxonomic classification and degree of novelty (novel species, genus, etc) of the recovered bins were obtained from the Microbial Genome Atlas (MiGA) webserver (http://microbial-genomes.org). Bin abundance was determined as the total length of all matching metagenomic or metatranscriptomic reads to the binned contigs from BLASTn searches (identity >=98% and fraction of read aligned >= 50%) divided by the metagenomic or metatranscriptomic sample sizes (in millions of reads) and the length of the bin genomes in Kbp (Kilo base pairs).

Phylogenetic reconstruction of bins was performed based on the concatenated alignment of universal single-copy proteins identified for each bin using the “HMM.essential.rb” script of the enveomics collection (20). For this, thirty-one bacterial and nine archaeal proteins present in the corresponding bins were extracted and multiple alignments for each protein were generated using ClustalΩ. Concatenated alignments without invariable sites were generated for archaeal and bacterial alignments using the script “Aln.cat.rb”. Phylogenetic reconstructions were determined using in RAxML v8.0.19 (-f a, -m PROTGAMMAAUTO, –N 100) and visualized in iTol.

N cycle protein sequences in the co-assembly and bins were detected using hidden Markov models obtained from FUNGENE (21), using HMMer (22). Detected target N cycle proteins were manually curated, when necessary, by assessing the presence of characteristic amino acid and phylogenetic congruency.

5.3.7 Phylogenetic Trees and Placement of Short-reads

To assess the phylogenetic affiliation of metagenomic or metatranscriptomic reads, reference and fully assembled protein sequences were aligned using ClustalΩ (23) with default parameters. Resulting alignments were used to build phylogenetic trees in RAxML v8.0.19 (24). Short-reads encoding the protein of interest were extracted from metagenomes or metatranscriptomes using ROCker (BLASTx) and placed in their corresponding phylogenetic tree using the methodology previously described (10).

Quantification of the number of reads assigned to a specific clade (e.g., to distinguish between *nxrA* or *narG* reads) was done using the "JPlace.distances.rb" script, also available in the enveomics collection.

To quantify *nirK* gene fragments assigned to specific clades, the same process as described above was repeated except that all reads detected by multiple ROCker models to previously described clades (25) (clades I+II, III and *Thaumarchaeotea*) were used.

5.3.8 Shotgun Metaproteomics

Approximately 10 gr of soil was collected from the 8 day control and ^{15}N-NH_4^+ amended microcosms and stored at $^{-}80°C$. Approximately 2 gr of soil from each sample was immersed in dodecyl sulfate lysis buffer and boiled for 15 minutes. Proteins were extracted using trichloroacetic acid precipitation. Extracted peptides were subjected to 24h multi-step chromatographic separation using an HPLC system (Ultimate 3000) connected to a mass spectrometer. Fragmented peptides were detected using a LTQ-Orbitrap mass spectrometer based on the multi-dimensional protein identification technology (MuDPIT) approach. For protein identification, raw spectra of two technical replicates from control and N-amended soils were searched against a custom database. This database included protein sequences predicted from the soil metagenome assemblies and reference proteomes for the 47 most common soil organisms (See Appendix D, Table D.5), and were used in the Myrimatch v2.1 algorithm. Average normalized spectral counts (Avg Spc) were determined for two run replicates as previously proposed (26) and used for quantification purposes.

5.4 RESULTS

5.4.1 Nitrification Activity in Soil Microcosms

The addition of an ammonium and urea solution (40 mM NH_4^+ and 20 mM urea) to soils was followed by the disappearance of ammonium, and nitrate accumulation over time, indicating nitrification activity (Figure 5.1). After a lag phase in activity during the first incubation points (1-2 days), a high rate of nitrification (~2 μNO_3^--N g^{-1} soil d^{-1}), measured as nitrate accumulation, was observed after 2 days. Accumulation of nitrate increased from an initial value of 0.58 ±0.3 μNO_3^--N g^{-1} soil to 15.9 ± 3.2 μNO_3^--N g^{-1} soil after 8 days of incubation. Ammonia concentration peaked after 24h of incubation (18.8 ± 0.42 $\mu g\text{-}NH_4^+$ -N g^{-1} soil) and decreased to 4.4 ± 2.6 $\mu g\text{-}NH_4^+$-N g^{-1} soil after 8 days of incubation. Headspace $^{15}N\text{-}N_2O$ flux derived from the measurement of stable nitrogen isotope (derived from biotic or abiotic $^{15}N\text{-}NH_4^+$oxidation) was detected after 24h of incubation (1.5 ± 0.4 $pg^{15}N\text{-}N_2O$ g^{-1} h^{-1}) increasing to 21.2 ±9.0 $^{15}N\text{-}N_2O$ pg^{-1} h^{-1} after 8 days of incubation (Figure 5.1). As expected, incubations receiving only irrigation water (i.e., no N amendment) did not show ammonia oxidation as initial NH_4^+ values ranged from 0.75 $\mu g\text{-}NH_4^+$ -N g^{-1} at the beginning of the incubation to 1.86 $\mu g\text{-}NH_4^+$-N g^{-1} after 8 days and no N_2O was detected. No significant changes in moisture content were observed (values ranged from 7.71% to 8.49% across microcosms and time points), and pH decreased in the N-amended samples after 8 days of incubation (from 7.53 to 6.3), which was consistent across replicated incubations (Appendix D, Table D.1).

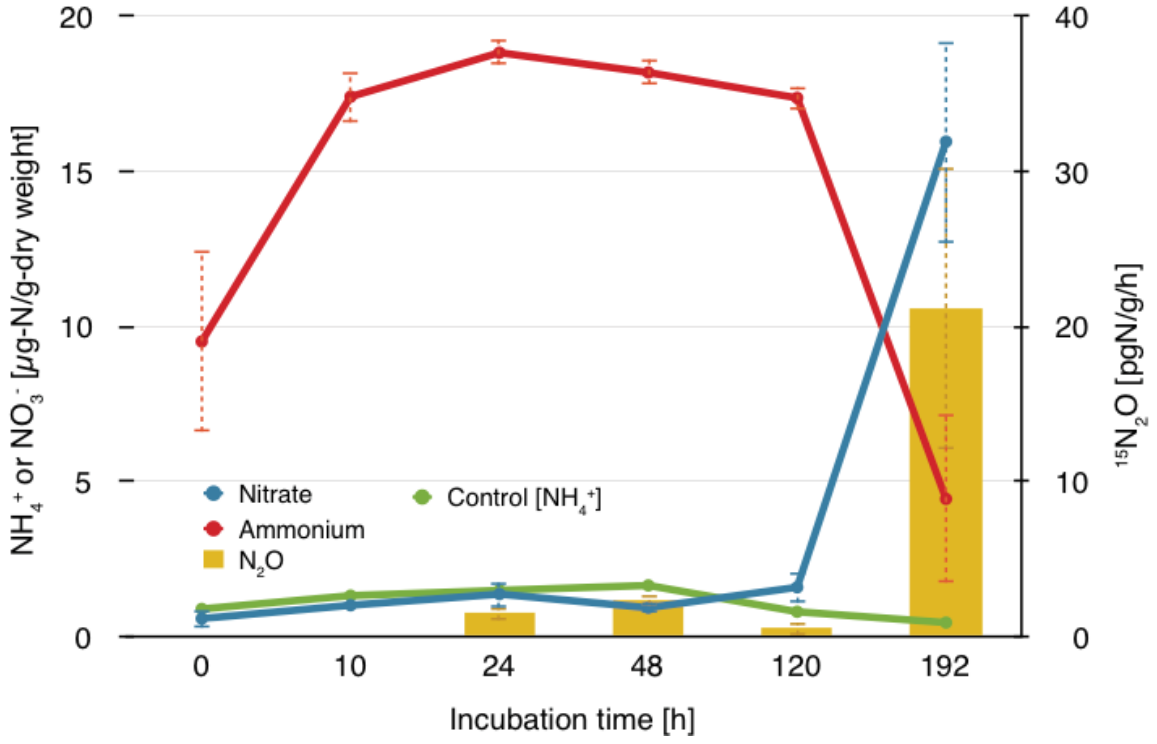


Figure 5.1. Nitrification activity in soil incubations amended with NH_4^+ and urea

Mean NH_4^+ and NO_3^- concentrations were determined at each incubation time point and used to determine total nitrification activity. Error bars represent the standard deviation (n=3).

5.4.2 Soil Metagenomes and Metatranscriptomes

A total of 13 soil shotgun Illumina metagenomes and 7 metatranscriptomes were obtained from control and amended soil incubations. Metagenomes ranged from 8.2 to 53.4 and metatranscriptomes from 10.1 to 31.3 million short-reads per sample (Appendix D, Tables D.2 and D.3). The estimated average coverage based on read redundancy (16) ranged from 0.27 to 0.49 for the soil metagenomes. The co-assembly of selected soil metagenomes generated 1.13 million contigs over 500 bp (assembly N50=1,206) and 2.07 million predicted genes.

A high fraction of ribosomal RNA was detected for all metatranscriptomes ranging from 94% to 98% of the total sequences (Appendix D, Table D.4). As expected based on the length of the rRNA genes, 23S rRNA/16S rRNA ratios ranged from 1.7 to 2.1, indicating an adequate quality RNA. Bacterial rRNA was the most abundant as indicated by its 16S rRNA abundance ranging from 30.2% to 35.9% of total transcripts per sample. Archaeal and eukaryotic 16S rRNA and 18S rRNA were less abundant, with values ranging from 0.09% to 0.15% and 0.53 to 2.9%, respectively.

5.4.3 Microbial Soil Populations at the 16S rRNA Gene Level

The examination of the taxonomic composition using recovered 16S rRNA (16S) gene and transcripts from N-amended incubations showed generally stable abundances for main microbial groups after 10h, 24h, 48h, 5 days, and 8 days of incubation. Metagenomes showed that, at the class taxonomical level,

Actinobacteria, *Gammaproteobacteria*, and *Alphaproteobacteria* were the most abundant groups accounting for more than 51% of the total community in N-amended incubation sets (Appendix D, Figure D.1). On the other hand, taxonomic composition derived from metatranscriptomes was also stable during the incubations but showed different abundances for main taxonomical groups compared to metagenomes. For instance, *Betaproteobacteria*, *Gammaproteobacteria*, and *Flavobacteria* were among the most abundant groups in cDNA samples, accounting for, on average, 66.2% of the total abundance. In agreement with our previous results from the same agricultural site, bacterial and archaeal groups associated to nitrification processes were comparatively less abundant than the aforementioned groups at both DNA and cDNA datasets. For instance, known AOB and NOB genera such as *Nitrosomonas* and *Nitrospira* had an average relative abundance 0.01% and 1.5% of the total population in metagenomes, respectively. Additionally, AOA genera belonging to *Nitrososphaera* and *Nitrosopumilus* showed an average relative abundance of 1.6% and 0.3% in metagenomes, respectively. Similar relatively low abundances were determined for known nitrifier genera in metatranscriptomes, but conspicuous increases in 16S transcript abundances were observed for bacterial nitrifier groups during the incubation period (Appendix D, Figure D.2a). For instance, when relative expression ratios (cDNA/DNA) were compared at the 16S level, AOB and NOB belonging to *Nitrosomonas* and *Nitrospira* showed over 6-fold increase between 10h and 8days incubation points. In contrast to these bacteria, archaeal groups

Nitrososphaera and *Nitrosopumilus* showed stable expression ratios and a peak in relative expression at 48h of incubation (Appendix D, Figure D.2a).

5.4.4 Individual Populations from Incubation Metagenomes

The assembly and binning of the soil metagenomes recovered 48 populations (bins) mostly representing *Proteobacteria* (41.6%), *Actinobacteria* (14.6%) and *Bacteroidetes* (10.4%) phyla. Most the recovered bins likely represented novel genera (22/48) and species (19/22), with the remaining bins likely representing new families (4/48) and subspecies (3/48), when the taxonomic novelty was evaluated using 7,373 reference genomes in MiGA (Appendix D, Table D.6) using genome-aggregate amino acid identity (AAI) thresholds for taxonomic rank delineation (27). Two recovered *Nitrospira* bins shared 61.48% average amino acid identity (AAI) (SD: 19.62%, based on 1793 shared proteins) indicating their similarity at the genus level (i.e., *Nitrospira*) (28). We also included two additional *Nitrospira* bins (bin021 and bin017) and four *Thaumarchaeota* bins representing the lineages I.1b (bin032 and bin019) and I.1a (bin109 and bin001) obtained from a previous analysis of field samples from the same site as the soil inoculum used in our soil incubations in the present study (10). These bins were likely missed in this study due to comparatively lower sequencing effort achieved or because of sample heterogeneity (e.g. lower abundance), but had relatively higher abundance in the previous field samples. Interestingly, *Nitrospira* bin007 and bin043 were closely related to previously described soil commamox (e.g., bin017) organisms sharing 68.07% AAI (SD: 17.87% based on 2263 shared proteins) and 72.13% AAI (SD: 24%, based on

1719 shared proteins), respectively. However, the two *Nitrospira* bins (007 and 043) lacked *amoA* genes and only bin043 encoded a full-length nitrite oxidoreductase gene (*nxrA*). In fact, these *Nitrospira* bins formed an independent cluster but close to the soil Comammox organisms, when reconstructed phylogenies using concatenated single-copy genes were evaluated (Appendix D, Figure D.3). Thus, the AAI values support the affiliation of bin007 and bin043 to *Nitrospira*, but the lack of genes involved in nitrification (probably due to low sequencing effort) might indicate their divergence from previously described soil Comammox organisms.

Genome-average relative expression ratios (cDNA/DNA) showed that some of the nitrifiers bins belonging to *Nitrospira* and *Thaumarchaeota* increased their activity throughout the incubation. For instance, *Thaumarchaeota* bin028 (I.1a) and bin019 (I.1b) showed maximum expression ratios changes of 5.1- and 2.5-fold after 5 days of incubation (compared to expression levels at 10h incubation). On the other hand, *Nitrospira* bin017 and bin007 showed the highest changes in expression values for bacterial nitrifiers of 2.1- and 3.4-fold also after 5 days of incubation (Appendix D, Figure D.2b)

5.4.5 Quantification of Nitrification Genes and Metagenomic Populations in Microcosms

To explore the microbial activity participating in nitrification processes in incubated soils, we specifically quantified gene fragments and transcripts directly involved in nitrification reactions. Relative expression ratios (cDNA/DNA) belonging to the urease subunit c (*ureC*) showed stable gene expression

throughout the incubation but relatively low average abundances compared to other nitrification genes. The relative expression (cDNA/DNA) of the bacterial ammonia monooxygenase subunit alpha (*amoA*) showed a conspicuous increase of over 12.5-fold compared to the expression values at 10h of incubation (Figure 5.2a). A further assessment of the phylogenetic affiliation of the *amoA* transcripts (cDNA) showed that detected fragments were mostly affiliated with *Betaproteobacteria* and showed up to 66-fold increase between some of the time points (Figure 5.2a). Unlike betaproteobacterial *amoA*, transcripts belonging to Comammox *Nitrospira* showed a stable abundance across incubations. A qualitative different pattern was obtained from *amoA* metagenomic reads (DNA level), where gene fragments belonging to *Nitrospira* (i.e., Comammox) were more abundant compared to *Betaproteobacteria amoA* reads until 5 days of incubation (Figure 5.2b). However, after 8 days of incubation, betaproteobacterial *amoA* DNA abundance increased 66-fold, whereas Comammox *amoA* gene fragments remained stable. The latter results indicate that Comammox may be more abundant under field conditions, which was also consistent with our previous study (10). Even though the relative expression for the archaeal *amoA* was more stable throughout the incubation compared to its bacterial counterpart, at 5 days of incubation, a ~5.3-fold increase compared to the previous sampling point was observed, indicating that archaeal AmoA activity temporarily increased at later time points of the incubation. Archaeal *amoA* transcripts belonging to the group I.1b were ~7 times more abundant than their I.1a counterpart across the incubations (Figure 5.2d). Similar to bacterial *amoA* patterns, betaproteobacterial

hydroxylamine oxidoreductase (*hao;* $NH_2OH \rightarrow NO_2^-$) also showed steady increase in relative expression across the incubation time points. Increasing *hao* transcripts were affiliated to *Betaproteobacteria* (up to 96% at 8 days of incubation) whereas Comammox *hao* fragments showed a stable abundance throughout the incubation, similar to the *amoA* patterns mentioned above (Figure 5.2c). As expected, bacterial nitrite oxidoreductase subunit alpha (*nxrA;* $NO_2^- \rightarrow NO_3^-$) expression also showed increased values, up to 12.4-fold, compared to 10h time point, consistent with the patterns observed for the previous nitrification genes.

Unexpectedly, *nirK* ($NO_2^- \rightarrow NO$) affiliated to *Thaumarchaeota* showed increased expression values compared to *nirK* fragments assigned to the *Nitrospira* clade (i.e., clade III). In fact, throughout the incubation, *Thaumarchaeota nirK* showed 8.5-fold increase after 8 days of incubation, indicating that *Thaumarchaeota* might have been more active in the reduction of nitrite compared to other steps of nitrification. Similar to *amoA*, transcripts belonging to the clade 1.b showed a 3.4-fold increase from 10hr to 8 days whereas reads from clade 1.a showed a stable abundance throughout the incubations (Figure 5.2e).

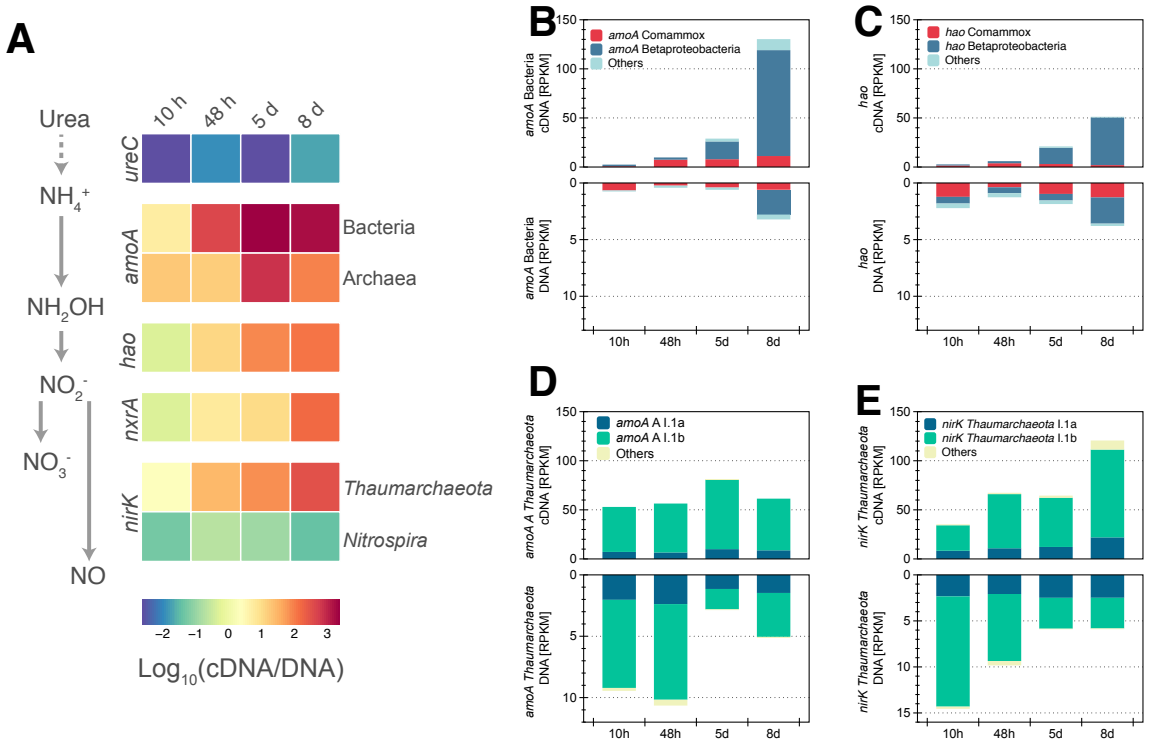


Figure 5.2. Nitrification genes in incubated soils

A. Relative expression ratios for each nitrification step in incubated soils were determined for 10h, 48h, 5d and 8 d of incubation. B and C show determined RPKM values for bacterial *amoA* (B) and *hao* (C) transcripts determined metatranscriptomes (top values) and gene fragments from metagenomes (bottom values). D and E show determined values for *Thaumarchaeota amoA* (D) and *nirK* transcripts determined from metatranscriptomes and gene fragments metagenomes.

In summary, it appeared as if *Betaproteobacteria* nitrifier populations, but not bacterial Comammox, responded first and rapidly to the N amendment based on the metatranscriptomic profiles, with their archaeal *Thaumarchaeota* counterparts following with less evident transcriptome shifts. The bacterial response, and to a lesser extent the thaumarchaeotal one, was also reflected at the DNA level, albeit with substantial delay in time, e.g., shifts were observed during the first couple days at the transcript level vs. 5-8 days since the start of incubation for the DNA level. These data were consistent across the individual nitrification steps and genes, with the probable exception of the archaeal *nirK*, and indicated that at least the *betaproteobacterial* nitrifiers grew substantially in response to N addition.

5.4.6 A Proteomic Perspective

A metaproteomic analysis of the control and N-amended samples at 8 days of incubation recovered a total of 2,892 and 1,629 non-redundant peptides, respectively. 844 peptides were detected in both incubation points whereas 2,048 and 785 were exclusively present in the control and N-amended samples, respectively. The top 20 most abundant proteins in the control samples were related to housekeeping and transport proteins whereas in the N-amended, oxidoreductases for small carbon and alcohol molecules and ATP synthesis were among the most abundant proteins detected (Appendix D, Table D.7). The taxonomic composition derived from annotated peptides showed 2.2-fold increased abundance for *Nitrospira* peptides when the number of unique peptides assigned to proteins was used as a proxy for abundance after

normalizing for the total number of peptides detected and averaging detected peptide counts for individual proteins (Figure 5.3a). The annotation of detected peptides in N-amended soils, using SEED functional categories, showed that protein folding and synthesis categories were among the most abundant and had similar abundances between the two treatments. However increased ATP synthases and transcription were observed in the N-amended samples relative to the control, presumably as a consequence of a higher microbial activity generated after the N input. On the other hand, heat-shock and degradation proteins were more abundant in un-amended (control) soil samples, probably reflecting a more prevailing dormant state for the microbial communities in these samples (Figure 5.3b). However, unlike metagenomes and metatranscriptomes datasets, a smaller fraction of peptides involved in nitrification were detected. For instance, the only detected peptides directly involved in nitrification pathways corresponded to the nitrite oxidoreductase subunit B (NxrB), which showed a 31.3% abundance increase in the N-amended samples compared to the control.

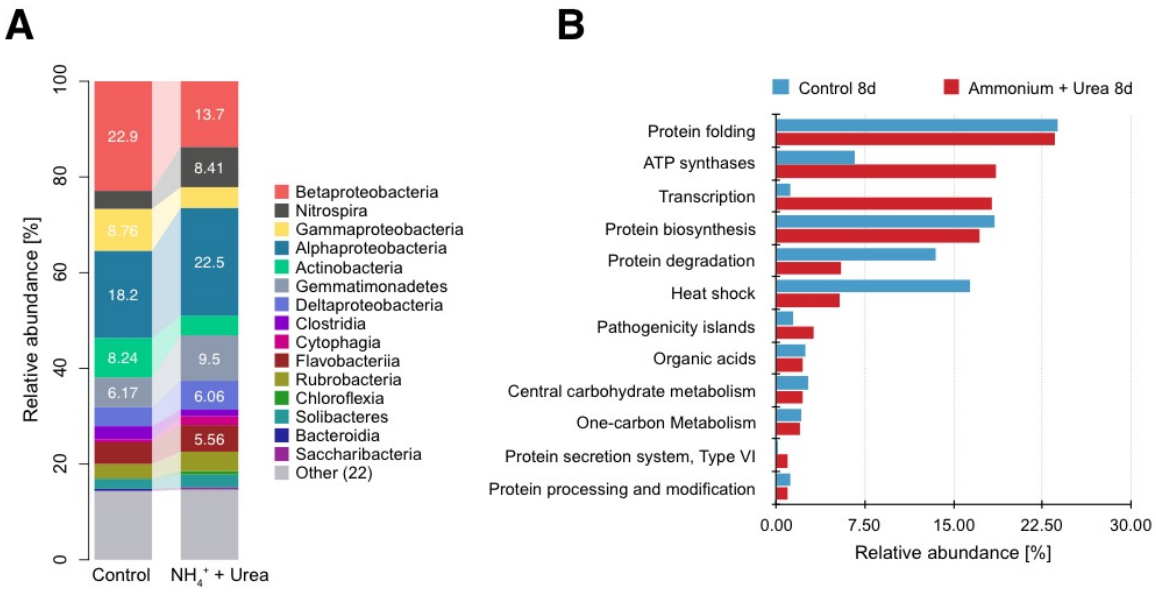


Figure 5.3 Metaproteomic analyses of incubated soils at 8 days of incubation

Panel A shows taxonomic affiliation (class) and abundance (average spectral counts) for peptides detected in control and N-amended incubations. Panel B shows summarized functional annotation of detected peptides using SEED functional categories.

5.4.7 Multi-omic Data as Proxies of Microbial Activity

The quantification of gene and transcripts involved in nitrification allowed us to analyze their relationship with measured nitrification rates in amended soils. Significant positive linear regressions were observed for bacterial *amoA* gene (genome equivalent) and transcript fragments (RPKM) with nitrate accumulation (Figure 5.4a). In addition, bacterial *amoA* and *nxrA* transcripts showed higher R^2 values, i.e., >0.95, compared to their gene counterparts, i.e., ranging from ~0 to 0.72, when regressions were made using measured nitrate or N_2O accumulation rates (Figure 5.4b). Other genes directly involved in nitrification pathways such as the archaeal *amoA* did not show significant relationships between measured nitrate and measured N_2O and gene or transcript abundances. Transcript abundances of other genes participating in nitrification or denitrification such as *hao* and *nosZ* were less predictive than bacterial *amoA* and *nxrA* transcripts, e.g., showed R^2 values below 0.3

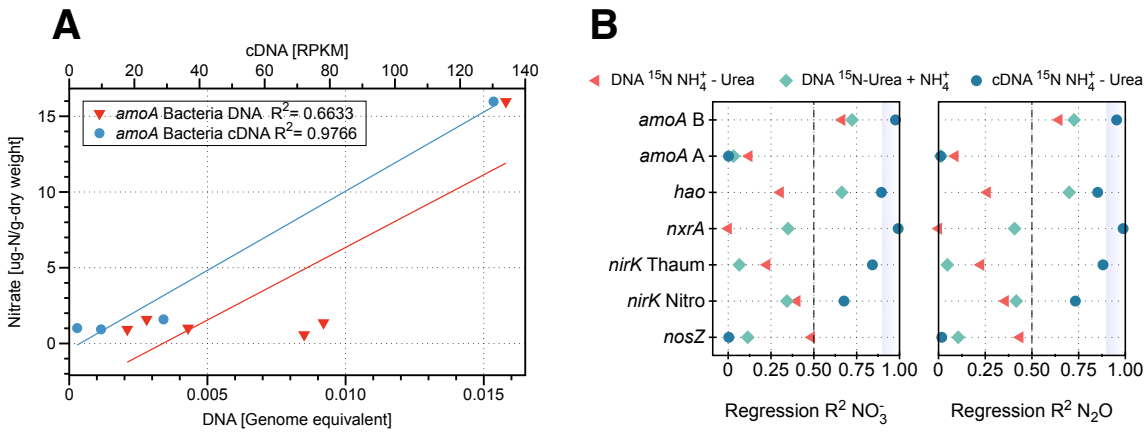


Figure 5.4. Regression analyses using metagenomes and metatranscriptomes as predictors of microbial activity

(A) Relationships between bacterial and archaeal *amoA* using metagenomes and metatranscriptomes and nitrate accumulation production. (B) Summary of determined R^2 values from linear regressions between nitrate accumulation or N_2O production and nitrification markers. Shapes indicate values obtained from gene and transcript abundances.

5.5 DISCUSSION

5.5.1 Using Multi-omic Approaches for Examining Measured Process Rates

Measuring nitrification process rates in incubated soils allowed us to evaluate the predictive power of high-throughput omic approaches in a highly diverse soil system. Even though all three omic approaches revealed increased abundance for target genes, transcripts and proteins related to nitrification processes, they showed differences in temporal resolution and quantitative capabilities. For instance, metatranscriptomic data showed the strongest correlation to the observed nitrification processes (i.e., ammonia or nitrite oxidation) within the first couple days of incubation points, whereas metagenomes only reflected the ongoing nitrification process only about after eight days of incubation (e.g., Figure 5.2). These data were presumably attributed to the fact that growth (e.g., at least a few replication cycles) should occur before metagenomics can reveal shifts over time. Therefore, metagenomics could also reflect active processes if the processes are ongoing for some time and are coupled with the growth of the corresponding organisms.

On the other hand, the obtained metaproteomes offered mostly a qualitative glimpse at nitrification processes and were less definitive in identifying common nitrification markers. The latter could be explained by the low number of detected proteins compared to the number of metagenomic and metatranscriptomics reads recovered that encoded the proteins of interest. Even though regression models using metatranscriptomes showed a better fit to measured nitrification processes (e.g., bacterial *amoA* or *nxrA* transcripts), our

datasets included two additional time points (t=0 and 24 h) for metagenomic datasets (Figure 5.4a). The exclusion of these values from the regression calculations improved the fit of DNA gene abundances to measured process rates. However, metatranscriptomes were still better at reflecting microbial activity for nitrification processes during all incubation points. Although more frequent sampling and incubations under different conditions will be required for more robust conclusions to emerge on the exact relationship(s) between molecular level information and process rates, the results reported here provided a first view of this relationship for soils, and are highly promising for the future.

Previous reports have also found metatranscriptomic approaches as better predictors of measured microbial activity (7) in controlled laboratory systems amended with exogenous organic compounds, but have been more limited in providing insights into the whole-microbial community response to the amendment. For instance, the changes in transcripts observed at early incubation points for specific lineages (e.g., Comammox vs. *Betaproteobacteria amoA*) reflected ongoing microbial activity (growth) that became only evident at the last incubation point in metagenomes (Figure 5.2b). Future incubation studies could shed further light on the intrinsic differences between nitrifier communities by testing variables such as oxygen availability (i.e., water saturation) and different agricultural soil types. Also, the use of nitrification inhibitors can help elucidate the origin of the measured N_2O for which our data is limited in predicting whether it has an biotic or abiotic origin. Thus, the integration of *in-situ* rates along with the microbial dynamics examined by metatranscriptomes and

metagenomes could provide the means to better understand and predict nitrification and N_2O emission in agricultural soils.

5.5.2 New Insights for Nitrification Pathways Derived from Agricultural Soil Microbial Communities

The metagenomic and metatranscriptomic datasets combined with phylogenetic approaches provided a closer examination of the poorly studied microbial diversity in agricultural soils. For instance, betaproteobacterial *amoA* genes have been commonly assayed as a proxy for bacterial ammonia oxidation in soils but this is not the case for genes and transcripts affiliated to the recently described Comammox *Nitrospira* (29, 30). Our results showed that even though *Betaproteobacteria amoA* transcripts responded to the addition of ammonium and urea, the relative abundance for *amoA* Comammox transcripts were stable (i.e., not responding to the N amendment) and Comammox populations were relatively more abundant in metagenomes in field samples. This observation agrees with our previous metagenomic results from the same agricultural field, where Comammox *amoA* genes and the organisms encoding these genes represented the highest fraction of the AOB communities (10). The differences between measured genes and transcripts indicated that the incubation conditions favored the activity of *Betaproteobacteria* over Comammox nitrifying bacteria, suggesting ecophysiological differences among these taxa for the incubation conditions or added substrates compared to field conditions.

The sequencing of isolates and environmental AOA genomes has shown

that even though they encode an AmoA protein, they lack a canonical hydroxylamine oxidation pathway (31). Recent findings have proposed that nitric oxide is essential for hydroxylamine oxidation to nitrite in archaea (32). The proposed mechanism first oxidizes ammonium to hydroxylamine followed by a consecutive oxidation to nitrite catalyzed by a putative Cu-protein that uses nitric oxide as an electron source. Interestingly, nitric oxide has been proposed to be derived from the activity of the NirK enzyme present in all AOA sequenced genomes. Our results show that unlike AOA *amoA* or canonical bacterial *nirK* transcripts, *Thaumarchaeota nirK* transcripts exhibited an increased abundance in the incubated soils, supporting the abovementioned hypothesis. Therefore, even though AOA *amoA* transcripts did not show clear changes in abundances compared to their bacterial counterpart, these results might be in agreement with previous hypothesis, and likely denote an unaccounted role *for Thaumarchaeota nirK* in nitrification in agricultural soils.

5.5.3 Multi-omic Limitations

From the technical point of view, soil samples are challenging to analyze not only because of their heterogeneous structure and chemical composition but also because of the highly diverse dwelling microbial communities and slow growth kinetics, in general. Despite the advancements presented here, there are still opportunities for further improvement. For instance, here we analyzed total RNA extractions from soils where ribosomal rRNA transcripts represented 94-98% of the total sample, limiting our study to a small fraction of transcripts related to functional genes. Current experimental approaches offer successful rRNA

depletion, when RNA yields are not limiting, for environmental samples (33). Additionally, all the results represented here only provide relative abundances for measured microbial markers. For instance, approaches such as qPCR or internal standards spiked into the DNA or cDNA library for sequencing (33) can strengthen and provide absolute quantification compared to those presented here.

Proteomic approaches offered a third layer of information for the microbial activity, but it was less comprehensive compared to metagenomes and metatranscriptomes in our study. Even though our database for matching detected spectra included a high fraction of nitrification proteins predicted from these agricultural soils, only peptides belonging to the nitrite oxidoreductase were detected. These results were presumably attributable, at least partially, to the low biomass, especially for the low abundance nitrifiers targeted here. Further, possible protein extraction biases due to the complexity of soil matrices as well as limited extraction of membrane proteins, such AmoA, might have also influenced the outcome of our efforts (34). Nonetheless, several peptides belonging to housekeeping proteins of nitrifier organisms showed an increased abundance during the incubation time, consistent with the results from metagenomic and metatranscriptomic approaches. Therefore, metaproteomics provided at least a qualitative confirmation of the underlying nitrification processes ongoing during our incubations and the responding taxa. Alternative proteomic approaches focused on a preselected set of proteins (i.e., selected reaction monitoring or target proteomics) could be used to explore low abundant

nitrification proteins. For instance, targeted proteomic approaches have been successfully used to study proteins in low abundance involved in bioremediation pathways in highly-diverse environmental systems (35).Therefore, targeted proteomics might offer new opportunities for researchers interested in detecting low-abundance peptides and prediction of process rates in complex samples (36).

The analyses of different omic techniques obtained from these incubations showed high correspondence between nitrification gene markers and nitrification process rates. The gene fragments and transcripts were mostly affiliated to novel nitrifier populations similar to those previously described in field soil metagenomes from the same agricultural site (10). The combination of metagenomic and metatranscriptomic approaches used here provided an promising strategy for examining microbial activity in agricultural soil environments. Therefore, the findings presented here highlighted the potential of omics data to serve as reliable proxies for examining microbial processes *in situ*, especially in soils, which has been proven to be among the most challenging tasks for environmental studies.

.

5.6 REFERENCES

1. **Ravishankara AR**, **Daniel JS**, **Portmann RW**. 2009. Nitrous oxide (N_2O): the dominant ozone-depleting substance emitted in the 21st century. Science **326**:123–125.

2. **Kool DM**, **Dolfing J**, **Wrage N**, **Van Groenigen JW**. 2011. Nitrifier denitrification as a distinct and significant source of nitrous oxide from soil. Soil Biology and Biochemistry **43**:174–178.

3. **Prosser JI**, **Nicol GW**. 2012. Archaeal and bacterial ammonia-oxidisers in soil: the quest for niche specialisation and differentiation. Trends in Microbiology **20**:523–531.

4. **Gubry-Rangin C**, **Nicol GW**, **Prosser JI**. 2010. Archaea rather than bacteria control nitrification in two agricultural acidic soils. FEMS Microbiology Ecology **74**:566–574.

5. **Hettich RL**, **Sharma R**, **Chourey K**, **Giannone RJ**. 2012. Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. Current Opinion in Microbiology **15**:373–380.

6. **Hug LA**, **Thomas BC**, **Sharon I**, **Brown CT**, **Sharma R**, **Hettich RL**, **Wilkins MJ**, **Williams KH**, **Singh A**, **Banfield JF**. 2015. Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. Environmental Microbiology **18**:159–173.

7. **Helbling DE**, **Ackermann M**, **Fenner K**, **Kohler H-PE**, **Johnson DR**. 2012. The activity level of a microbial community function can be predicted from its metatranscriptome. The ISME Journal **6**:902–904.

8. **Schneider T**, **Keiblinger KM**, **Schmid E**, **Sterflinger-Gleixner K**, **Ellersdorfer G**, **Roschitzki B**, **Richter A**, **Eberl L**, **Zechmeister-Boltenstern S**, **Riedel K**. 2012. Who is who in litter decomposition|[quest]| Metaproteomics reveals major microbial players and their biogeochemical functions. The ISME Journal **6**:1749–1762.

9. **Hultman J**, **Waldrop MP**, **Mackelprang R**, **David MM**, **McFarland J**, **Blazewicz SJ**, **Harden J**, **Turetsky MR**, **McGuire AD**, **Shah MB**, **VerBerkmoes NC**, **Lee LH**, **Mavrommatis K**, **Jansson JK**. 2015. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature **521**:208–212.

10. **Orellana LH**, **Chee-Sanford JC**, **Sanford RA**, **Löffler FE**, **Konstantinidis KT**. Year-round metagenomes reveal stable microbial communities in agricultural soils and novel ammonia oxidizers responding to fertilization. Submitted.

11. **Yang WH**, **Traut BH**, **Silver WL**. 2015. Microbially mediated nitrogen retention and loss in a salt marsh soil. Ecosphere **6**:art7.

12. **Yang WH**, **Teh YA**, **Silver WL**. 2011. A test of a field-based ^{15}N–nitrous oxide pool dilution technique to measure gross N_2O production in soil. Global Change Biology **17**:3577–3588.

13. **Orellana LH**, **Rodriguez-R LM**, **Higgins S**, **Chee-Sanford JC**, **Sanford RA**, **Ritalahti KM**, **Löffler FE**, **Konstantinidis KT**. 2014. Detecting nitrous oxide reductase (NosZ) genes in soil metagenomes: method development and implications for the nitrogen cycle. mBio **5**:e01193–14.

14. **Cox MP**, **Peterson DA**, **Biggs PJ**. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics **11**:485.

15. **Zhang J**, **Kobert K**, **Flouri T**, **Stamatakis A**. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics **30**:614–620.

16. **Rodriguez-R LM**, **Konstantinidis KT**. 2014. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics **30**:629–635.

17. **Kopylova E**, **Noé L**, **Touzet H**. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics **28**:3211–3217.

18. **Wang Q**, **Garrity GM**, **Tiedje JM**, **Cole JR**. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol **73**:5261–5267.

19. **Orellana LH**, **Rodriguez-R LM**, **Konstantinidis KT**. 2017. ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. Nucleic Acids Res **45**:e14.

20. **Rodriguez-R LM**, **Konstantinidis KT**. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints **4**:e1900v1.

21. **Fish JA**, **Chai B**, **Wang Q**, **Sun Y**, **Brown CT**, **Tiedje JM**, **Cole JR**. 2013. FunGene: the functional gene pipeline and repository. Frontiers in Microbiology **4**:291.

22. **Eddy SR**. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol **7**:e1002195.

23. **Sievers F**, **Wilm A**, **Dineen D**, **Gibson TJ**, **Karplus K**, **Li W**, **Lopez R**, **McWilliam H**, **Remmert M**, **Söding J**, **Thompson JD**, **Higgins DG**. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol **7**:539–539.

24. **Stamatakis A**. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688–2690.

25. **Wei W**, **Isobe K**, **Nishizawa T**, **Zhu L**, **Shiratori Y**, **Ohte N**, **Koba K**, **Otsuka S**, **Senoo K**. 2015. Higher diversity and abundance of denitrifying microorganisms in environments than considered previously. The ISME Journal **9**:1–12.

26. **Paoletti AC**, **Parmely TJ**, **Tomomori-Sato C**, **Sato S**, **Zhu D**, **Conaway RC**, **Conaway JW**, **Florens L**, **Washburn MP**. 2006. Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. PNAS **103**:18928–18933.

27. **Luo C**, **Rodriguez-R LM**, **Konstantinidis KT**. 2014. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. Nucleic Acids Res **42**:e73–e73.

28. **Konstantinidis KT**. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci USA **102**:2567–2572.

29. **Daims H**, **Lebedeva EV**, **Pjevac P**, **Han P**, **Herbold C**, **Albertsen M**, **Jehmlich N**, **Palatinszky M**, **Vierheilig J**, **Bulaev A**, **Kirkegaard RH**, **Bergen von M**, **Rattei T**, **Bendinger B**, **Nielsen PH**, **Wagner M**. 2015. Complete nitrification by Nitrospira bacteria. Nature **528**:504–509.

30. **van Kessel MAHJ**, **Speth DR**, **Albertsen M**, **Nielsen PH**, **Op den Camp HJM**, **Kartal B**, **Jetten MSM**, **Lücker S**. 2015. Complete nitrification by a single microorganism. Nature **528**:555–559.

31. **Stahl DA**, **la Torre de JR**. 2012. Physiology and Diversity of Ammonia-Oxidizing Archaea. Annu Rev Microbiol **66**:83–101.

32. **Kozlowski JA**, **Stieglmeier M**, **Schleper C**, **Klotz MG**, **Stein LY**. 2016. Pathways and key intermediates required for obligate aerobic ammonia-dependent chemolithotrophy in bacteria and Thaumarchaeota. The ISME Journal 1–10.

33. **Tsementzi D**, **Poretsky R**, **Rodriguez-R LM**, **Luo C**, **Konstantinidis KT**. 2014. Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. Environ Microbiol Rep **6**:640–655.

34. **VerBerkmoes NC**, **Denef VJ**, **Hettich RL**, **Banfield JF**. 2009. Systems Biology: Functional analysis of natural microbial consortia using community proteomics. Nature Reviews Microbiology **7**:196–205.

35. **Werner JJ**, **Ptak AC**, **Rahm BG**, **Zhang S**, **Richardson RE**. 2009. Absolute quantification of Dehalococcoides proteins: enzyme bioindicators of chlorinated ethene dehalorespiration. Environmental Microbiology **11**:2687–2697.

36. **Hood LE**, **Omenn GS**, **Moritz RL**, **Aebersold R**, **Yamamoto KR**, **Amos M**, **Hunter Cevera J**, **Locascio L**. 2012. New and improved proteomics technologies for understanding complex biological systems: Addressing a grand challenge in the life sciences. PROTEOMICS **12**:2773–2783.

CHAPTER 6. CONCLUSIONS AND RECOMMENDATIONS

Soil dwelling-microorganisms play a key role in maintaining nutrient cycles, which sustain all forms of life. By using a combination of genomic, transcriptomic and proteomic approaches, this thesis contributed to the advancement of current experimental and bioinformatic techniques that can improve the examination of microorganisms in the environment.

At the moment this thesis began, little was known about how to accurately examine target genes in short-read metagenomes. This is a major challenge of modern sequencing technologies since most search algorithms used to annotate short-reads are designed to accurately work with full-length protein sequences. In chapter 2, we introduced a new bioinformatic approach to solve this problem, called ROCker, which determines precise parameters for detecting gene fragments related to target sequences of interest from any metagenomic or metatranscriptomic dataset. The use of ROCker proved to be clearly advantageous compared to the common practice of selecting arbitrary filtering parameters. Future challenges for this area will be the improvement of databases in order to ensure that reference sequences can truly reflect the latest biochemical findings of target protein enzymes. One alternative approach for obtaining well curated gene references is the use of databases such as FunGene (1) that reflects the joint effort of experts and researchers for the annotation of key microbial proteins. In addition, ROCker relies on the simulation of complex metagenomic datasets using an algorithm called GRINDER (2). Even though this

method reliably simulates short-read metagenomes and their associated sequencing errors, upcoming sequencing technologies might require different simulation parameters not currently offered in GRINDER. Furthermore, the idea underlying ROCker can also be extended to full-length sequence searches, thus, having broad applications in bioinformatic sequence analysis.

Having established an accurate and reliable method to study target genes in metagenomes, we studied nitrous oxide genes, *nosZ,* in different soil and marine sediments, which was described in the third chapter. The discovery of functional atypical NosZ proteins has opened the possibility that a much larger number of microorganisms than previously expected with unaccounted N_2O-reducing potential contribute to lowering the N_2O emissions into the atmosphere. The abundance and diversity of atypical *nosZ* genes were likely missed in previous PCR-based surveys because typical *nosZ* sequences were used for primer design. This issue underscores a limitation of PCR-based methods when studying target genes in the environment. Similar to *nosZ* genes, much of the information currently used for designing PCR primers comes from available bacterial isolates. These microorganisms do not necessarily represent well the natural microbial diversity found in the environment, thus, biasing the PCR efforts for identifying and quantifying genes in the environment. One recommendation is to use comprehensive approaches, such as metagenomics, in order to unveil the natural sequence diversity of target genes, which can then be used to design and generate PCR primers for more accurate gene quantifications. Current research focuses on understanding enzyme kinetics (3) and the effect of agricultural

management in the distribution of *nosZ*-harboring microbial populations (4), emphasizing the importance of the these population in soil ecosystems. In fact, the inclusion of these microbial populations in predictive models might help future research aiming to reduce the impact of microbes in the generation of greenhouse gases (e.g., N_2O emissions) from engineered or agricultural systems.

Agricultural sites receiving an excess of nitrogen fertilizer are among the largest anthropogenic sources of nitrous oxide, a potent greenhouse gas. In the fourth chapter of this thesis, we examined the microbial communities associated with nitrification and denitrification in two agricultural sites with a long history of agricultural usage. The use of metagenomic approaches facilitated the discovery of novel archaeal (*Thaumarchaeota*) and bacterial (Comammox *Nitrospira*) communities that showed clear responses to the addition of nitrogen fertilizers. In addition, these results provided evidence that archaeal organisms might respond to N fertilization much more dramatically than expected for their presumed oligotrophic nature. Therefore, it appears that the agricultural management in the two soils has selected discrete-evolving nitrifier microorganisms differing from canonical nitrifiers. Future studies and models assessing the impact of microorganisms in the nitrogen cycle should include the novel archaeal and bacterial diversity presented here. In addition, the results presented here can be used for the accurate gene and transcript quantification (e.g., quantitative PCR) to be combined with measured process rates (e.g., nitrification) during a nitrogen fertilization event in the agricultural field. Collectively, the methodologies

presented here offer an approach to accurately target and track natural microbial communities involved in key nutrient cycles and processes in the environment.

The results presented in chapters 2, 3 and 4, showcased the strength of metagenomic approaches for studying target genes and metagenomic bin populations in soils. However, the study of the DNA sequences only reflects the genetic potential found in a particular sample and not necessarily microbial activity. In chapter 5, we tested the power of omic techniques to predict the microbial activity in nitrogen amended soils measured biochemically based on either the quantification of gene (DNA level), transcripts (mRNA level), and peptides. The high correlation observed between transcript related to ammonia oxidation and N_2O or nitrate accumulation offered a proxy for examining microbial activity in soils. This approach can be employed in natural or engineered systems in order to identify molecular markers to be used as a proxy for measuring microbial activity. For instance, in agricultural systems, the assessment of the microbial activity combined with field data measurements can be used to accurately define the microbial impact on N_2O emissions during the growing season in agricultural sites. Despite the advancements presented here, there are still opportunities for further improvements. Even though the soils were incubated under well aerated conditions, the high soil heterogeneity characteristics might still produce low oxygen microenvironments where denitrification could proceed. In order to distinguish N_2O production by nitrifiers and denitrifiers, the use acetylene prevents generation of N_2O through nitrification by inactivating the bacterial and archaeal AmoA activity (5, 6). In order to further strengthen these

findings, different incubation conditions such as variation in the oxygen availability and different N mass and source (e.g., NO_3^-) input will improve the generation of robust predictive models for natural nitrifier communities in soils. In addition, much more limited results were obtained from proteomic analyses: only one of the key nitrification markers (NxrB) was detected when a shot-gun proteomic approach was tested. However, nitrifiers (and their proteins) are known for being less abundant than other soils microorganisms, which was also reflected by our metagenomic and transcriptomic results. Thus, future studies interested in detecting and describing peptides from low abundance organisms should either improve the protein extraction method to enrich low abundance peptides (7) or use approaches targeted at specific proteins. In addition, in this chapter we analyzed total RNA extractions from soils where rRNA represented 94-98% of the total RNA sample, limiting our study to a small fraction of non-ribosomal gene transcripts related to functional genes. Current experimental approaches offer a high success for rRNA depletion using environmental samples (8) but generally at the expense of a higher mass of total extracted RNA required, sometimes limited in low biomass soil samples. Thus, future studies should use RNA extraction methods that allow high RNA quantities without losing quality. Therefore, novel experimental approaches using these recommendations can be designed to advance our understating of *in situ* microorganisms catalyzing the cycling of key nutrients in natural ecosystems.

6.1 REFERENCES

1. **Fish JA**, **Chai B**, **Wang Q**, **Sun Y**, **Brown CT**, **Tiedje JM**, **Cole JR**. 2013. FunGene: the functional gene pipeline and repository. Frontiers in Microbiology **4**:291.

2. **Angly FE**, **Willner D**, **Rohwer F**, **Hugenholtz P**, **Tyson GW**. 2012. Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Res **40**:e94.

3. **Berg P**, **Klemedtsson L**, **Rosswall T**. 1982. Inhibitory effect of low partial pressures of acetylene on nitrification. Soil Biology and Biochemistry **14**:301–303.

4. **Offre P**, **Prosser JI**, **Nicol GW**. 2009. Growth of ammonia-oxidizing archaea in soil microcosms is inhibited by acetylene. FEMS Microbiology Ecology **70**:99–108.

5. **Siciliano SD**, **Ma WK**, **Ferguson S**, **Farrell RE**. 2009. Nitrifier dominance of Arctic soil nitrous oxide emissions arises due to fungal competition with denitrifiers for nitrate. Soil Biology and Biochemistry **41**:1104–1110.

6. **Keiblinger KM**, **Wilhartitz IC**, **Schneider T**, **Roschitzki B**, **Schmid E**, **Eberl L**, **Riedel K**, **Zechmeister-Boltenstern S**. 2012. Soil metaproteomics – Comparative evaluation of protein extraction protocols. Soil Biology and Biochemistry **54**:14–24.

7. **Nesatyy VJ**, **Suter MJF**. 2007. Proteomics for the Analysis of Environmental Stress Responses in Organisms. Environmental science & technology **41**:6891–6900.

8. **Tsementzi D**, **Poretsky R**, **Rodriguez-R LM**, **Luo C**, **Konstantinidis KT**. 2014. Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. Environ Microbiol Rep **6**:640–655.

APPENDIX A: SUPPLEMENTARY MATERIAL FOR CHAPTER 2

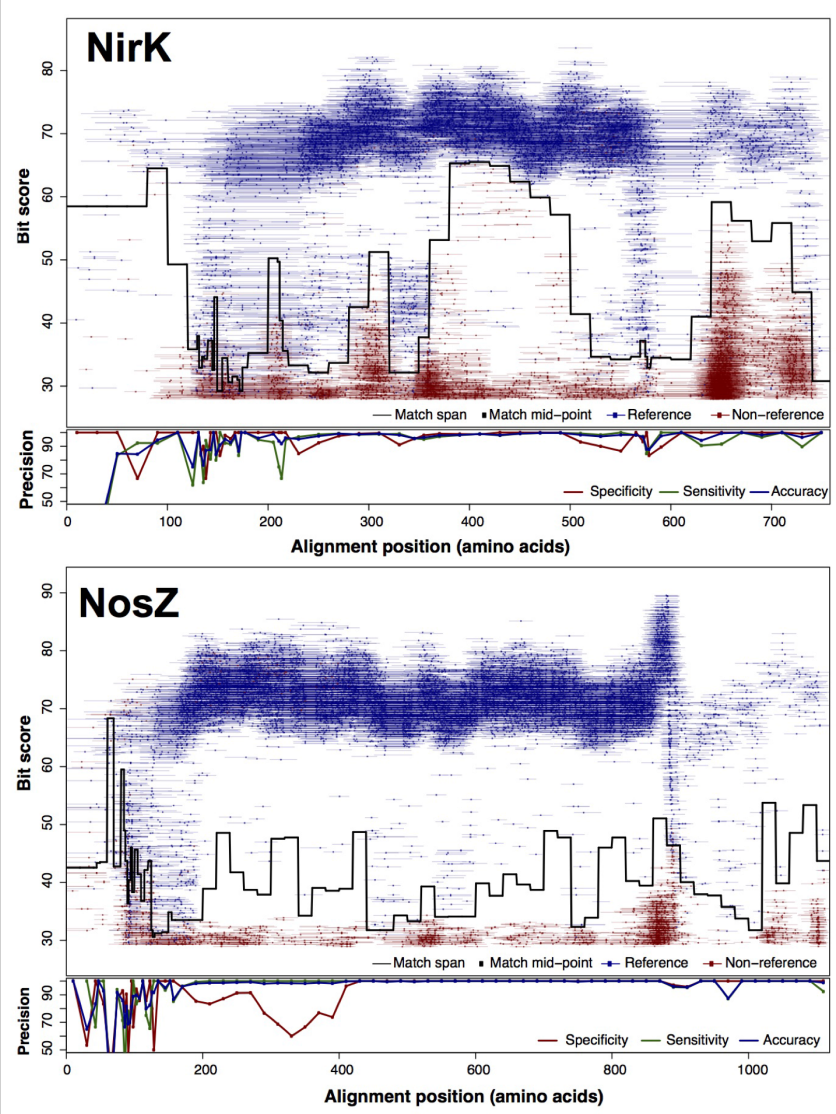


Figure A.1. Best-discriminating thresholds calculated for NirK and NosZ reference protein sequences using ROCker. The top panels display the bitscore (y-axis) of the matches from simulated datasets (100 bp reads) encoding the target gene, i.e., derived from the reference sequences (true positive; blue) or a non-target sequence (false positive; red). Each matching read is represented as a line based on the coordinates of the alignment against the reference sequences where the read maps to, and the dot represents the middle point of the read. The traversing black line represents the calculated ROCker best

bitscores for consecutive windows of variable length. The bottom panel presents summary statistics on the performance of each window based on true and false positives (TP and FP, respectively) and true and false negatives (TN and FN, respectively). In particular, specificity (TN/[FP+TN]; red), sensitivity (TP/[TP+FN]; green), and accuracy ([TP+TN]/[TP+FP+TN+FN]; blue) are shown. Note that the summary statistics for the entire alignment may differ from the arithmetic mean across windows because the number of matches is not uniformly distributed along the alignment. See Material and Methods section for more details.

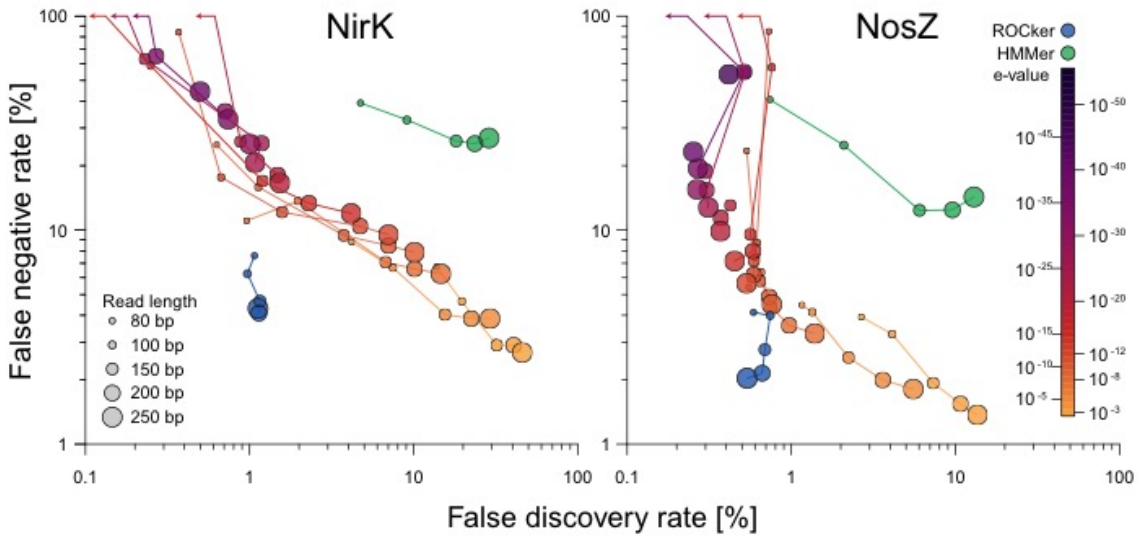


Figure A.2. Comparison of false negative and false positive rates for simulated datasets of different read lengths using ROCker profiles and e-value thresholds. Simulated shotgun datasets of 80, 100, 150, 200, and 250 bp read length were generated using ROCker and searched against reference protein sequences with DIAMOND. The outputs were filtered using the calculated ROCker profiles and fixed e-value thresholds, similar to Figure 1.

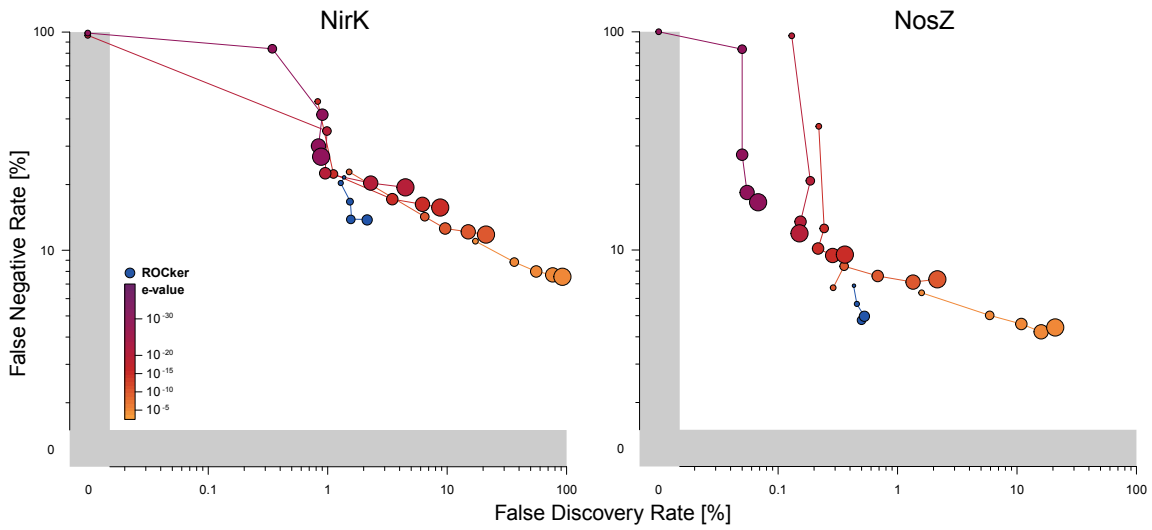


Figure A.3. False negative and false positive rates for a tenfold cross-validation test for ROCker profiles and e-value thresholds. Simulated datasets of 100,150,200, 250 and 300 bp read lengths were generated using ROCker following a tenfold cross-validation strategy (see Materials and Methods for details), and searched against reference proteins sequences using BLASTx. The similarity search outputs were filtered using ROCker profiles and fixed e-values, similar to Figure 1.

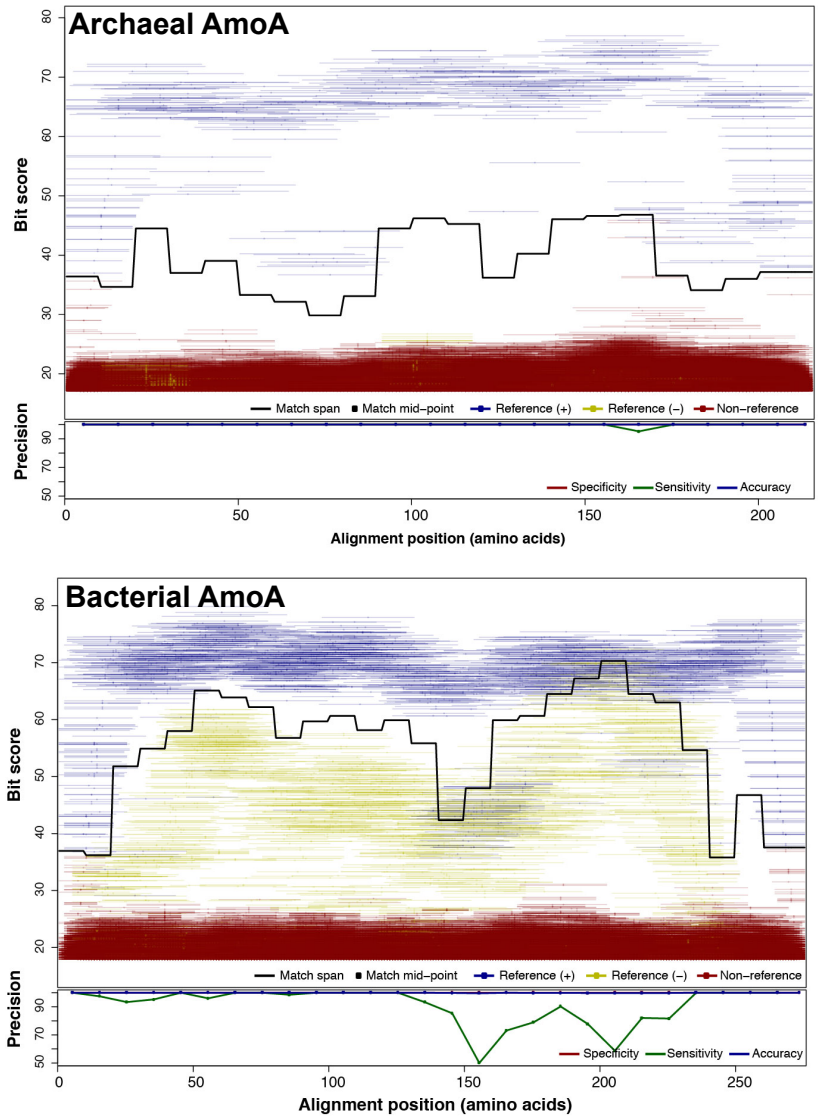


Figure A.4. Best-discriminating thresholds calculated for bacterial and archaea AmoA reference protein sequences using ROCker. The top panel displays the bitscore (y-axis) of the matches from simulated datasets (100 bp reads) encoding the target gene, i.e., derived from the reference sequences (true positive; blue), a non-target sequence (false positive; red), and negative references (false positive; gold). Each matching read is represented as a line based on the coordinates of the alignment against the reference sequences where the read maps to, and the dot represents the middle point of the read. The traversing black line represents the calculated ROCker best bitscores for consecutive windows of variable length. The bottom panel presents summary statistics on the performance of each window based on true and false positives (TP and FP, respectively) and true and false negatives (TN and FN, respectively). In particular, specificity (TN/[FP+TN]; red), sensitivity (TP/[TP+FN]; green), and accuracy ([TP+TN]/[TP+FP+TN+FN]; blue) are shown. Note that the

summary statistics for the entire alignment may differ from the arithmetic mean across windows because the number of matches is not uniformly distributed along the alignment. See Material and Methods section for more details.

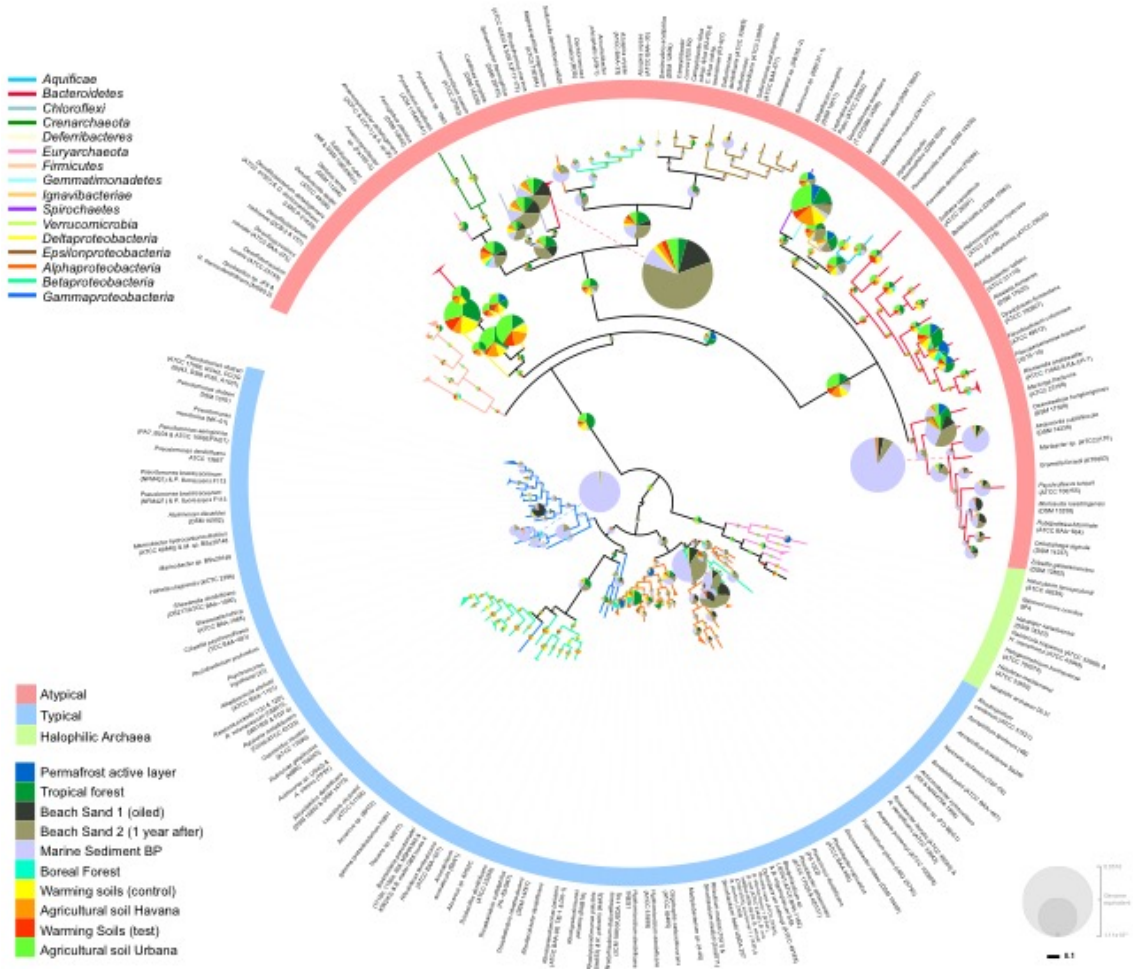


Figure A.5. Maximum likelihood NosZ phylogenetic reconstruction of *nosZ* reads recovered from terrestrial and marine metagenomes. 173 sequences extracted from sequenced bacterial and archaeal genomes were manually verified as NosZ and used for the construction of the reference NosZ phylogenetic tree. ROCker-filtered and translated metagenomic reads from 10 metagenomes were placed in the tree. The radii of the pie charts represent the abundance for each node (calculated as genome equivalents).

Table A.1. Simulated datasets generated by ROCker for NosZ, NirK, AmoA and RpoB reference sequences.

Reference	Length	Total reads	References	Non-references
NosZ	80	21,341,581	11,658	21,329,923
	100	21,341,581	11,672	21,329,909
	150	21,341,581	12,132	21,329,449
	200	21,341,581	12,386	21,329,195
	250	21,341,581	12,841	21,328,740
NirK	80	17,805,483	6,729	17,798,754
	100	17,805,483	6,856	17,798,627
	150	17,805,483	7,150	17,798,333
	200	17,805,483	7,391	17,798,092
	250	17,805,483	7,586	17,797,897
AmoA archaea + negative references	80	5,558,303	324	5,557,979
	100	5,558,303	353	5,557,950
	150	5,558,303	320	5,557,983
	200	5,558,303	361	5,557,942
	250	5,558,303	382	5,557,921
AmoA bacteria + negative references	80	5,558,303	1,290	5,557,013
	100	5,558,303	1,310	5,556,993
	150	5,558,303	1,395	5,556,908
	200	5,558,303	1,511	5,556,792
	250	5,558,303	1,577	5,556,726
RpoB	80	8,306,471	9,727	8,296,744
	100	8,306,471	9,899	8,296,572

Table A.2. Statistics for BLASTx, DIAMOND, and HMMer for the NosZ and NirK simulated datasets.

		Time [minutes]			Memory [Gb]			FNR			FDR		
Dataset	**Length**	BLASTx	DIAMOND	HMMer	BLASTx	DIAMOND	HMMer	BLASTx	DIAMOND	HMMer	BLASTx	DIAMOND	HMMer
NirK	80	1,454	108	0.2	0.45	9.6	<1	7.25	7.59	39.26	0.92	1.07	4.74
	100	1,579	151	0.22	0.5	8.3	<1	6.04	6.26	32.76	0.89	0.97	9.09
	150	1,880	301	0.53	0.56	8.02	<1	4.59	4.63	26.1	0.92	1.16	18.12
	200	2,221	509	0.3	0.67	7.44	<1	4.29	4.09	25.35	1.06	1.14	23.55
	250	2,549	570	0.35	0.75	7.97	<1	4.28	4.28	26.9	1.08	1.13	28.74
NosZ	80	1,753	138	0.4	0.47	4.87	<1	4.35	4.13	40.75	0.56	0.59	0.74
	100	1,948	203	0.64	0.54	4.34	<1	3.85	3.98	24.91	0.84	0.74	2.09
	150	2,349	201	0.38	0.62	4.1	<1	2.32	2.77	12.37	0.81	0.69	6.04
	200	2,685	546	0.483	0.74	3.78	<1	2.34	2.14	12.44	0.62	0.66	9.57
	250	3,312	721	0.62	0.844	4.28	<1	1.59	2.03	14.28	0.69	0.54	12.96

Table A.3. Metadata for the short-read metagenomes used in this report.

Source	Name	Reads (trimmed)	Read length Q2 (trimmed)	Description	Location	Reference
Soils	Permafrost active layer	2,915,582	60	Core 1 active layer	Hess creek, Fairbanks (Alaska, USA)	Mackelprang et al., 2011
	Warming soils (Test)	135,285,168	78	Experimental soil (T5)	Kessler Farm Field Laboratory (Oklahoma, USA)	Luo et al., 2013
	Warming soils (Control)	170,759,503	80	Experimental soil (C5)		
	Havana	235,579,975	90	Agricultural Soil	Havana (Illinois, USA)	Orellana LH et al., 2014
	Urbana	347,895,127	100		Urbana (Illinois, USA)	
	Tropical forest (AR3)	3,979,301	89	Soil	Misiones (Argentina)	Fierer, N (2012)
	Marine sediment (oil spill)	248,713,907	101	Marine sediment	Gulf of Mexico (Florida, USA)	Mason, O (2014)
	Boreal forest (BZ1)	5,018,178	92	Soil	Bonanza creek (Alaska, USA)	Fierer, N (2012)
	Beach sand (oiled)	32,840,836	101	Contaminated beach sand	Pensacola beach (Florida, USA)	Rodríguez-R (2015)
	Beach sand (1 year after)	33,188,686	101			
Fresh water	Lake Lanier	24,534,098	100	Water (Lake)	Lake Lanier (Georgia, USA)	Rodríguez-R (2014)
Engineered	Waste water treatment	203,250,192	131	Waste water treatment plant	Odense (Denmark)	McIlroy, S (2014)
Human	Human Stool	8,311,833	82	Human Microbiome		Human Microbiome Project Consortium, 2012

Table A.4. Comparison between BLASTx and UProC for NosZ, NirK and the bacterial AmoA based on simulated datasets.

Simulated Datasets		UProC		BLASTx	
Target gene	Sequencing depth	FDR	FNR	FDR	FNR
AmoA Bacteria	1x	93.45	27.03	96.36	0.00
	5x	91.79	20.93	96.10	0.00
	10x	91.93	20.44	96.15	0.00
	20x	91.96	22.30	96.07	0.00
NirK	1x	87.86	38.71	87.33	0.31
	5x	88.48	42.49	87.36	0.57
	10x	88.31	41.62	87.25	0.64
	20x	88.34	42.13	87.30	1.20
NosZ	1x	80.03	45.13	81.13	0.39
	5x	79.92	44.67	81.07	0.42
	10x	80.02	44.70	81.21	0.37
	20x	80.05	44.70	81.08	0.16

Table A.5. Reconstruction of target references by Xander on simulated datasets.

Simulated dataset		References in dataset	Percentage of target references assembled			
			1X	5X	10X	20X
AmoA_bacteria		7	0.0	71.4 (1 FP)	128.6 (3 FP)	142.9 (3 FP)
NirK		134	0.0	23.9	49.3	52.2
NosZ	Typical	89	2.2	56.2	73.0	53.9
	Atypical	64	0.00	45.31	92.19	70.31

Table A.6. Comparison of ROCker and Xander using simulated datasets of different sequencing efforts. For Xander, FDR and FNR were computed using the mapping reads found in the program output for each target reference. For ROCker, FDR and FNR were determined as indicated in the text.

Simulated Dataset	Method	AmoA Bacteria		NirK		NosZ (Typical + Atypical)	
Sequencing depth		FDR	FNR	FDR	FNR	FDR	FNR
1x	ROCker	3.45	5.08	0.51	6.00	0.30	3.12
5x		2.03	8.95	1.22	5.77	0.50	3.23
10x		3.39	7.89	1.06	6.68	0.39	3.30
20x		2.51	9.70	0.95	6.49	0.30	3.33
1x	Xander	100.00	100.00	100.00	100.00	0.00	97.79
5x		32.34	43.17	0.33	77.98	0.16	67.76
10x		34.93	30.36	1.90	61.73	0.19	47.74
20x		30.11	10.79	1.66	57.65	0.25	46.39

APPENDIX B: SUPPLEMENTARY MATERIAL FOR CHAPTER 3

Supplementary Materials and Methods B.1.

Average coverage estimation for soil metagenomes

The average sequencing coverage obtained for each metagenome was estimated using the Nonpareil v2.2 algorithm (1). Briefly, forward (or single) reads of each metagenome were trimmed and input to Nonpareil. The abundance-weighted average coverage was estimated from the frequency of redundant reads (i.e., reads sharing >95% sequence identity) based on randomly drawn subsets of reads from each metagenome (Appendix B, Table B.5). Nonpareil curves were computed based on the redundancy values, and the sequencing effort necessary to achieve 95% average coverage was estimated by the projection of the curves as described (1).

Metagenome assembly and read annotation

De novo assembly of metagenomes was conducted using a hybrid protocol that combines Velvet (3), SOAPdenovo (4) and Newbler 2.2 (Roche) as previously described (5) (Table S5). Protein-coding reads were identified by FragGeneScan (6) using the Illumina 0.5% error model and then searched against the UniRef50 database (7), using the BLAT algorithm (8) with default settings (BLAST8 output). The best hit for each read, when longer than 17 amino acids, was used for further analysis. The number of best hits against the

archaeal, bacterial, viral and eukaryotic fractions of UniRef50 database were used as a proxy for calculating the relative abundance of each domain (Appendix B, Figure B.2).

Analysis of 16S rRNA gene fragments recovered in the metagenomes

Metagenomic reads encoding bacterial 16S rRNA gene fragments were detected by searching them against the GreenGenes (9) (October 2012 release) sequences clustered at the 79% nucleotide identity level using BLASTn (settings: no dust, word size 7, penalty -2, max target seqs 1, xdrop 150 and e-value cut-off 0.001). Reads with alignments of at least 50 bp in length and 70% sequence identity were considered to encode 16S rRNA genes and were extracted from the metagenomes. These 16S rRNA gene-encoding reads were subsequently searched against a 97% nucleotide sequence identity clustered version of GreenGenes for taxonomic assignment according to their best match when better than 97% sequence identity. The number of sister reads assigned to the same taxon (e.g., phylum or class) was normalized by dividing by the total number of 16S rRNA gene-encoding reads assigned to a taxon in each metagenome and used as a proxy of the taxon relative abundance (Fig. S3). Single-reads or paired reads that matched different taxa were not used in this analysis.

REFERENCES

1. **Rodriguez-R LM, Konstantinidis KT**. 2013. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics. doi: 10.1093/bioinformatics/btt584

2. **Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert J a, Wall DH, Caporaso JG**. 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc. Natl. Acad. Sci. U. S. A. **109**:21390–5.

3. **Zerbino DR, Birney E**. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–9.

4. **Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J**. 2010. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. **20**:265–72.

5. **Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT**. 2012. Individual genome assembly from complex community short-read metagenomic datasets. ISME J. **6**:898–901

6. **Rho M, Tang H, Ye Y**. 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. **38**:e191. doi: 10.1093/nar/gkq747

7. **Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH**. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics **23**:1282–8.

8. **Kent WJ**. 2002. BLAT---The BLAST-Like Alignment Tool. Genome Res. **12**:656–664.

9. **DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL**. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. **72**:5069–72.

Supplementary Results B.1.

Description of the soil metagenomes

From the two composite DNA samples representing the microbial communities in the Havana sand and Urbana silt loam agricultural soils, a total of 384 and 398 million pair-end reads were obtained after sequencing, respectively. Only short contigs (N50 = 640 and 770 bp for Havana and Urbana soils, respectively; N50 represents the longest length for which the collection of all contigs of that length or longer contains at least half of the total of the lengths of the contigs), comprising a small fraction of the total reads (<5%) in either soil sample (Table S5), were obtained by *de novo* assembly. Similar N50 statistics to those obtained from both soil samples are expected for metagenomic datasets of low coverage and from highly diverse microbial communities (1). Therefore, our efforts to describe the abundance and diversity of *nosZ* genes were primarily focused on unassembled, quality-trimmed short-reads with an average length of 100 bp.

The average sequence coverage obtained (i.e., fraction of the genomes recovered in a sequencing dataset) was estimated to be two times greater for the sand than the silt loam metagenomes according to the Nonpareil algorithm (1). The calculated sequencing effort to cover 95% of the diversity of the communities in each sample was 0.4 and 1.35 Tbp for the sand and silt loam metagenomes, respectively. Bacteria represented the majority of the microbial communities in both agricultural soils based on taxonomic assignment of protein-encoding reads

(~93% of total), whereas the archaeal, eukaryotic, and viral reads represented minor fractions (~5% all together; Fig. S2). In addition, the taxonomic classification of 16S rRNA gene-encoding reads showed that the most abundant classes in the Havana sand included *Proteobacteria* (35%), *Acidobacteria* (19%), *Actinobacteria* (13%) and *Chlorofexi* (6%), whereas *Proteobacteria* (32%), *Actinobacteria* (17%), *Acidobacteria* (17%) and *Verrucomicrobia* (7%) were prominent in the Urbana silt loam soil (Fig. S3); similar to several other previously determined soil metagenomes [e.g., (2)].

REFERENCES

1. **Rodriguez-R LM, Konstantinidis KT**. 2013. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics. doi: 10.1093/bioinformatics/btt584

2. **Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert J a, Wall DH, Caporaso JG**. 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc. Natl. Acad. Sci. U. S. A. **109**:21390–5.

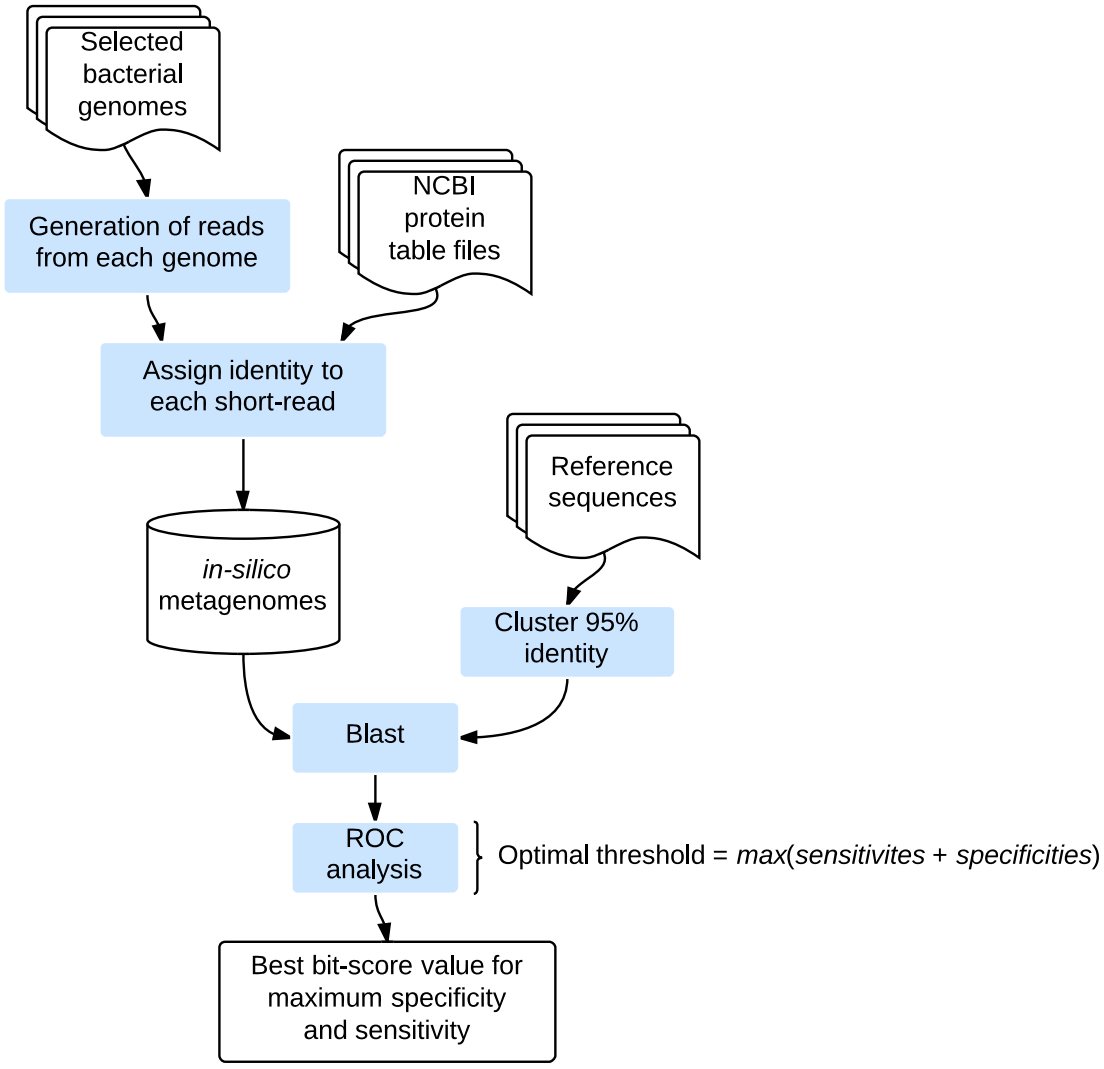


Figure B.1. Flow chart for calculating gene specific bitscore cut-offs.

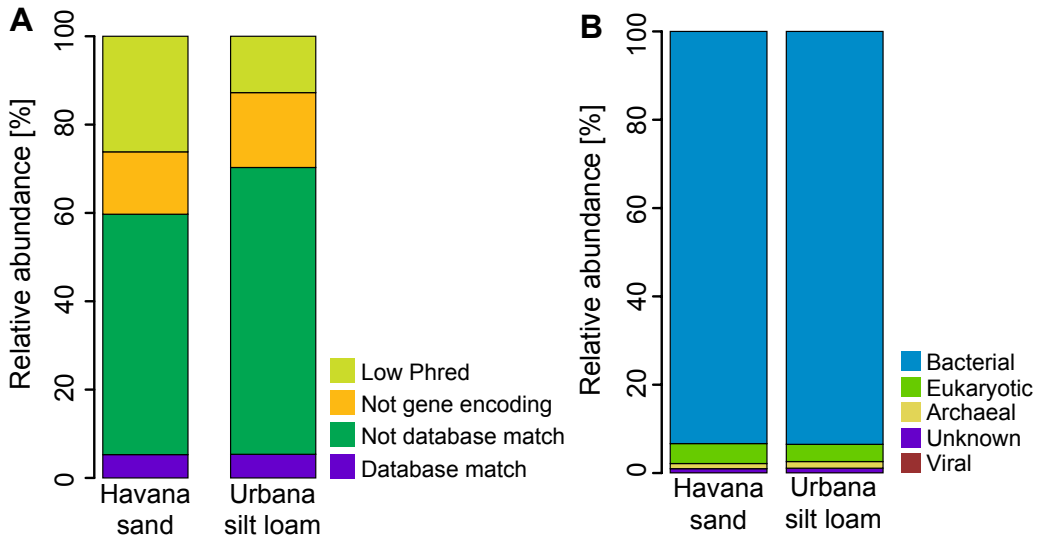


Figure B.2. Soil metagenomic library quality and domain-level composition based on taxonomic affiliation of protein-encoding reads. From the total amount of reads sequenced for each library, the stacked bars in A show the fraction of high quality, protein-encoding and reads matching the Uniref50 database. B shows the fraction of protein-encoding reads having a match in bacterial, eukaryotic and viral references sequences of the Uniref50 database.

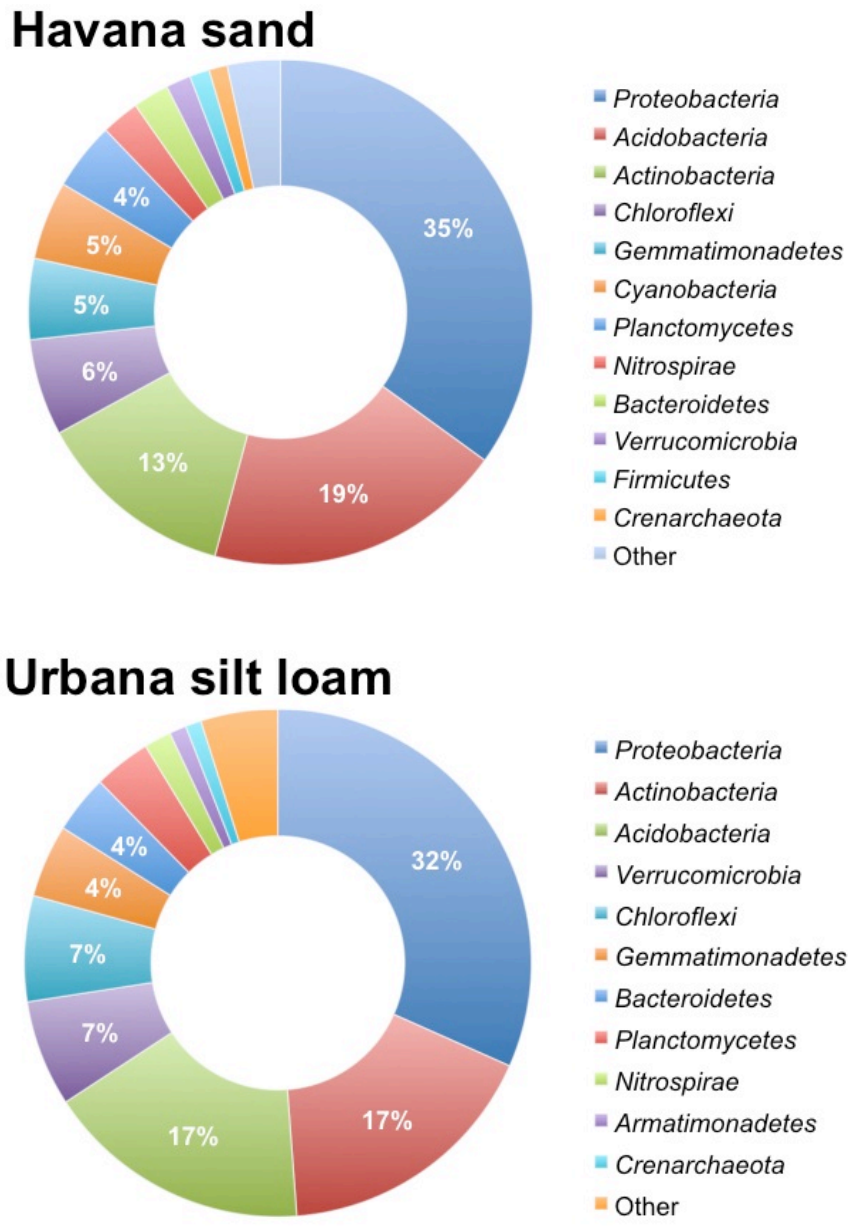


Figure B.3. **Taxonomic characterization of the metagenomes based on the recovered 16S rRNA gene reads.** Metagenomic reads encoding fragments of the 16S rRNA gene were extracted and searched against the GreenGenes database. The number of paired reads matching the same taxon was taken as proxy of taxon relative abundance and relative abundances are shown on the graph.

Table B.1. Genomes used to generate the *in silico* Library I and II. *Indicates the genomes whose housekeeping and *nosZ* genes were used for estimating the fraction of genomes encoding a *nosZ* gene in the agricultural metagenomes. Identifier (GI) numbers for both DNA molecules and NosZ protein are showed in the second and third column respectively.

Organism	Chromosome or secondary replicon	NosZ protein
Achromobacter xylosoxidans A8 *	311103224	311107725
Acidovorax ebreus TPSY	222109225	222110273
Acidovorax sp. JS42 *	121592436	121593552
Alicycliphilus denitrificans BC	319760738	319763751
Alkalilimnicola ehrlichii MLHE-1	114319166	114320233
Anaeromyxobacter dehalogenans 2CP-1	220915123	220916662
Anaeromyxobacter dehalogenans 2CP-C	86156430	86158824
Anaeromyxobacter sp. Fw109-5 *	153002879	152026694
Anaeromyxobacter sp. K	197120421	197121870
Aromatoleum aromaticum EbN1	56475432	56479025
Azoarcus sp. BH72	119896292	119899403
Azospirillum brasilense Sp245	392378906	392379406
Azospirillum lipoferum 4B	374998886	374999482

Bordetella petrii DSM 12804 *	163854304	163858659
Bradyrhizobium japonicum USDA 110 *	27375111	27375426
Bradyrhizobium sp. BTAi1	148251626	148257266
Brucella canis ATCC 23365	161620094	161620351
Brucella microti CCM 4915	256014795	256015059
Brucella ovis ATCC 25840	148557829	148558347
Brucella suis 1330 *	56968493	23500031
Burkholderia mallei ATCC 23344	53723370	53725328
Burkholderia pseudomallei 1106a	126451443	126454828
Burkholderia pseudomallei 1710b	76808520	76809000
Burkholderia pseudomallei 668	126438353	126438843
Burkholderia pseudomallei K96243	53717639	53719237
Burkholderia pseudomallei MSHR346	237810278	237812437
Burkholderia thailandensis E264	83718394	83718912
Caldilinea aerophila DSM 14535 = NBRC 104270	383760955	383761670
Campylobacter concisus 13826 *	157163852	157164517
Campylobacter curvus 525.92	154173617	154175149
Campylobacter fetus subsp. fetus 82-40	118474057	118475369

Candidatus Accumulibacter phosphatis clade IIA str. UW-1	257091663	257094967
Cellulophaga algicola DSM 14237	319951593	319954659
Colwellia psychrerythraea 34H	71277742	71281572
Cupriavidus metallidurans CH34	291481467	94313839
Dechloromonas aromatica RCB	71905642	71907203
Dechloromonas aromatica RCB *	71905642	71907207
Dechlorosoma suillum PS	372486701	372487564
Dechlorosoma suillum PS	372486701	372489916
Denitrovibrio acetiphilus DSM 12809	291285947	291286859
Desulfitobacterium dehalogenans ATCC 51507	392391692	392391859
Desulfitobacterium hafniense DCB-2	219666071	219666278
Desulfitobacterium hafniense Y51	89892746	89893007
Desulfomonile tiedjei DSM 6799	392408409	392409040
Desulfosporosinus meridiei DSM 13257	402570638	402573835
Desulfotomaculum ruminis DSM 2154	334338613	334340099
Dinoroseobacter shibae DFL 12	159042556	159045734
Dyadobacter fermentans DSM 18053 *	255033817	255034499
Ferroglobus placidus DSM 10642	288930407	288930531

Flavobacteriaceae bacterium 3519-10 *	255534169	255536235
gamma proteobacterium HdN1	304309652	304313365
Gemmatimonas aurantiaca T-27 *	226225406	226226791
Geobacillus thermodenitrificans NG80-2	138893679	138895390
Gramella forsetii KT0803 *	120434372	120435766
Hahella chejuensis KCTC 2396	83642913	83646128
Haliscomenobacter hydrossis DSM 1100	332661999	332667397
Haloarcula marismortui ATCC 43049	55376942	55377286
Halogeometricum borinquense DSM 11551	313117184	313117258
Halopiger xanaduensis SH-6	336252096	336253165
Halorubrum lacusprofundi ATCC 49239	222478439	222479592
Hydrogenobacter thermophilus TK-6	288817321	288817484
Hyphomicrobium denitrificans ATCC 51888	300021538	300023395
Ignavibacterium album JCM 16511 *	385808586	385809432
Leptospira biflexa serovar Patoc strain 'Patoc 1 (Ames)' *	189909570	189909940
Leptothrix cholodnii SP-6	171056692	171058167
Magnetospirillum magneticum AMB-1	83309099	83312185
Maribacter sp. HTCC2170	305664376	305665390

Marinobacter aquaeolei VT8	120552944	120555988
Marinobacter hydrocarbonoclasticus ATCC 49840	387812450	387815417
Marivirga tractuosa DSM 4126	313674129	313676798
Methylobacterium sp. 4-46	170738367	170741063
Neisseria lactamica 020-06	313667359	313668935
Nitratifractor salsuginis DSM 16511	319955760	319956251
Nitratiruptor sp. SB155-2	152989753	152991529
Ochrobactrum anthropi ATCC 49188	153010078	153011661
Oligotropha carboxidovorans OM5	209883160	209884173
Opitutus terrae PB90-1 *	182411826	182413621
Paracoccus denitrificans PD1222	119385557	119386924
Pedobacter saltans DSM 12145 *	325102757	325106344
Persephonella marina EX-H1	225849564	225850796
Photobacterium profundum SS9	54301680	54302526
Polymorphum gilvum SL003B-26A1	328541624	328542184
Prevotella denticola F0289	327312315	327312795
Pseudomonas aeruginosa LESB58	218888746	218890408
Pseudomonas aeruginosa PA7	152983466	152988445

Pseudomonas aeruginosa PAO1 *	110645304	15598588
Pseudomonas aeruginosa UCBPP-PA14	116048575	116051410
Pseudomonas brassicacearum subsp. brassicacearum NFM421	330806657	330808901
Pseudomonas brassicacearum subsp. brassicacearum NFM421	330806657	330809478
Pseudomonas fluorescens F113	378947941	378950908
Pseudomonas fluorescens F113	378947941	378951332
Pseudomonas mendocina NK-01	330500914	330501053
Pseudomonas stutzeri A1501 *	146280397	146283867
Pseudovibrio sp. FO-BEG1	374328350	374331474
Psychromonas ingrahamii 37	119943794	119945164
Pyrobaculum calidifontis JCM 11548	126458628	126460532
Pyrobaculum sp. 1860	374325525	374326246
Ralstonia eutropha H16	38637668	38637913
Ralstonia pickettii 12J *	187925958	187926244
Rhodobacter capsulatus SB 1003 *	294679055	294679128
Rhodobacter sphaeroides ATCC 17025	146279170	146279349
Rhodobacter sphaeroides KD131	221218192	221218312

Rhodoferax ferrireducens T118	89898822	89901968
Rhodopseudomonas palustris BisA53 *	115522030	115525101
Rhodopseudomonas palustris BisB18	90421528	90421952
Rhodopseudomonas palustris DX-1	316931396	316934764
Rhodopseudomonas palustris TIE-1	192288433	192290736
Rhodospirillum centenum SW	289546492	209967151
Rhodothermus marinus DSM 4252	268315578	268317562
Riemerella anatipestifer ATCC 11845 = DSM 15868	313205511	313207159
Robiginitalea biformata HTCC2501 *	260060589	260061020
Roseobacter denitrificans OCh 114 *	110677421	110678855
Ruegeria pomeroyi DSS-3	56708791	56708840
Salinibacter ruber DSM 13855	83814055	83816541
Salinibacter ruber M8	294505815	294506199
Shewanella denitrificans OS217	91791369	91793571
Shewanella loihica PV-4 *	127510935	127514328
Sinorhizobium meliloti 1021 *	16262453	16263096
Sphaerobacter thermophilus DSM 20745 *	269836033	269836766
Sulfurimonas autotrophica DSM 16294	307719921	307721903

Sulfurimonas denitrificans DSM 1251*	78776201	78777496
Sulfurimonas denitrificans DSM 1251	78776201	78777964
Sulfurovum sp. NBC37-1	152991597	152993755
Thauera sp. MZ1T *	237653092	217969098
Thermomicrobium roseum DSM 5159 *	221635406	221635600
Thioalkalivibrio sulfidophilus HL-EbGr7	220933193	220935728
Thiobacillus denitrificans ATCC 25259 *	74316018	74317407

Table B.2. **List of NosZ reference sequences representing more than 0.1% of total *nosZ* reads in soil metagenomes.** Representative NosZ sequences of the 95% clusters of NosZ sequences found exclusively in the sand or the silt loam soil metagenomes are highlighted in bold. NosZ representatives are alphabetically ordered according to typical followed by atypical classification.

Havana sand soil		Urbana silt loam soil	
NosZ representative	Relative abundance [%]	NosZ representative	Relative abundance [%]
Achromobacter xylosoxidans A8	0.33	*Achromobacter xylosoxidans* A8	0.18
***Acidovorax sp*. JS42**	0.83	*Alicycliphilus denitrificans* BC	0.24
Alicycliphilus denitrificans BC	1.09	***Alkalilimnicola ehrlichii* MLHE-1**	0.13
Aromatoleum aromaticum EbN1	1.52	*Aromatoleum aromaticum* EbN1	0.30
Azoarcus sp. BH72	1.38	*Azoarcus sp*. BH72	0.32
***Azospirillum brasilense* Sp245**	0.17	***Azospirillum lipoferum* 4B**	0.15
Bordetella petrii DSM 12804	0.38	*Bordetella petrii* DSM 12804	0.23
Bradyrhizobium japonicum USDA 110	1.23	*Bradyrhizobium* japonicum USDA 110	0.78
Bradyrhizobium sp. BTAi1	0.43	*Bradyrhizobium sp.* BTAi1	0.78
Brucella suis 1330	0.31	*Brucella suis* 1330	0.12
Burkholderia pseudomallei 668	0.59	*Burkholderia pseudomallei* 668	0.26

***Colwellia psychrerythraea* 34H**	0.12	*Cupriavidus metallidurans* CH34	0.12
Cupriavidus metallidurans CH34	1.14	*Hyphomicrobium denitrificans* ATCC 51888	0.26
***Dinoroseobacter shibae* DFL 12**	0.19	*Leptothrix cholodnii* SP-6	0.16
***gamma proteobacterium* HdN1**	0.31	***Marinobacter hydrocarbonoclasticus* ATCC 49840**	0.31
Hyphomicrobium denitrificans ATCC 51888	0.33	*Methylobacterium* sp. 4-46	0.39
Leptothrix cholodnii SP-6	1.42	***Neisseria lactamica* 020-06**	0.19
Methylobacterium sp. 4-46	0.64	*Ochrobactrum anthropi* ATCC 49188	0.19
Ochrobactrum anthropi ATCC 49188	0.85	*Oligotropha carboxidovorans* OM5	0.31
Oligotropha carboxidovorans OM5	0.83	*Paracoccus denitrificans* PD1222	0.22
Paracoccus denitrificans PD1222	0.64	*Polymorphum gilvum* SL003B-26A1	0.22
***Photobacterium profundum* SS9**	0.19	*Pseudomonas brassicacearum* subsp. brassicacearum NFM421	0.12
Polymorphum gilvum SL003B-26A1	0.33	***Pseudomonas stutzeri* A1501**	0.22
Pseudomonas brassicacearum subsp. brassicacearum NFM421	0.12	*Ralstonia eutropha* H16	0.11
Ralstonia eutropha H16	1.42	*Ralstonia pickettii* 12J	0.26
Ralstonia pickettii 12J	1.78	***Rhodobacter sphaeroides* ATCC 17025**	0.15

Rhodoferax ferrireducens T118	0.45	*Rhodoferax ferrireducens* T118	0.36
Rhodopseudomonas palustris BisA53	0.38	*Rhodopseudomonas palustris* BisA53	0.19
Rhodopseudomonas palustris BisB18	0.73	*Rhodopseudomonas palustris* BisB18	0.40
Rhodopseudomonas palustris DX-1	0.21	*Rhodopseudomonas palustris* DX-1	0.13
Rhodopseudomonas palustris TIE-1	0.38	*Rhodopseudomonas palustris* TIE-1	0.22
Rhodospirillum centenum SW	0.26	*Rhodospirillum centenum* SW	0.19
***Roseobacter denitrificans* OCh 114**	0.28	***Ruegeria pomeroyi* DSS-3**	0.11
Sinorhizobium meliloti 1021	1.07	*Sinorhizobium meliloti* 1021	0.31
Thauera sp. MZ1T	1.09	*Thauera* sp. MZ1T	0.47
Thiobacillus denitrificans ATCC 25259	3.06	***Thioalkalivibrio sulfidophilus* HL-EbGr7**	0.12
Anaeromyxobacter dehalogenans 2CP-C	5.22	*Thiobacillus denitrificans* ATCC 25259	0.38
Anaeromyxobacter sp. Fw109-5	7.47	*Anaeromyxobacter dehalogenans* 2CP-C	6.85
Caldilinea aerophila DSM 14535	1.40	*Anaeromyxobacter* sp. Fw109-5	8.31
Campylobacter concisus 13826	0.17	*Caldilinea aerophila* DSM 14535	1.59
Campylobacter fetus subsp. fetus 82-40	0.33	*Campylobacter concisus* 13826	0.12
Candidatus Accumulibacter	1.16	*Campylobacter fetus subsp. fetus*	0.18

phosphatis clade IIA		82-40	
Cellulophaga algicola DSM 14237	0.45	*Candidatus Accumulibacter phosphatis* clade IIA	0.22
Dechloromonas aromatica RCB	1.16	*Cellulophaga algicola* DSM 14237	0.47
Dechlorosoma suillum PS	2.11	*Dechloromonas aromatica* RCB	0.40
Dechlorosoma suillum PS	0.47	*Dechlorosoma suillum* PS	0.55
Denitrovibrio acetiphilus DSM 12809	0.28	*Dechlorosoma suillum* PS	0.36
Desulfitobacterium hafniense Y51	0.21	*Denitrovibrio acetiphilus* DSM 12809	0.27
Desulfomonile tiedjei DSM 6799	1.26	***Desulfitobacterium dehalogenans* ATCC 51507**	0.18
Desulfosporosinus meridiei DSM 13257	0.38	*Desulfitobacterium hafniense* Y51	0.34
Desulfotomaculum ruminis DSM 2154	0.69	*Desulfomonile tiedjei* DSM 6799	1.18
Dyadobacter fermentans DSM 18053	2.94	*Desulfosporosinus meridiei* DSM 13257	0.55
Ferroglobus placidus DSM 10642	1.33	*Desulfotomaculum ruminis* DSM 2154	1.39
Flavobacteriaceae bacterium 3519-10	2.73	*Dyadobacter fermentans* DSM 18053	6.15
Gemmatimonas aurantiaca T-27	5.31	*Ferroglobus placidus* DSM 10642	0.94
Geobacillus thermodenitrificans NG80-2	0.31	*Flavobacteriaceae bacterium* 3519-10	4.01
Gramella forsetii KT0803	1.40	*Gemmatimonas aurantiaca* T-27	5.96

Haliscomenobacter hydrossis DSM 1100	1.47	Geobacillus thermodenitrificans NG80-2	0.18
Hydrogenobacter thermophilus TK-6	3.84	*Gramella forsetii* KT0803	0.42
Ignavibacterium album JCM 16511	6.21	*Haliscomenobacter hydrossis* DSM 1100	3.30
Leptospira biflexa serovar Patoc strain	0.28	*Hydrogenobacter thermophilus* TK-6	7.40
Magnetospirillum magneticum AMB-1	0.31	*Ignavibacterium album* JCM 16511	5.79
Maribacter sp. HTCC2170	0.73	*Leptospira biflexa serovar Patoc* strain	0.40
Marivirga tractuosa DSM 4126	1.45	*Magnetospirillum magneticum* AMB-1	0.35
Nitratifractor salsuginis DSM 16511	0.12	*Maribacter sp.* HTCC2170	0.83
Opitutus terrae PB90-1	5.22	*Marivirga tractuosa* DSM 4126	1.00
Pedobacter saltans DSM 12145	2.77	*Nitratifractor salsuginis* DSM 16511	0.31
Persephonella marina EX-H1	3.11	*Opitutus terrae* PB90-1	11.35
Prevotella denticola F0289	0.57	*Pedobacter saltans* DSM 12145	2.30
Pyrobaculum sp. 1860	0.21	*Persephonella marina* EX-H1	3.78
Rhodothermus marinus DSM 4252	2.42	*Prevotella denticola* F0289	1.36
Riemerella anatipestifer ATCC 11845	2.42	***Pyrobaculum calidifontis* JCM 11548**	0.26

Robiginitalea biformata HTCC2501	0.47
Salinibacter ruber M8	0.38
Sphaerobacter thermophilus DSM 20745	1.21
Sulfurimonas denitrificans DSM 1251	0.40
Sulfurimonas denitrificans DSM 1251	0.14
Sulfurovum sp. NBC37-1	0.19
Thermomicrobium roseum DSM 5159	2.04

Pyrobaculum sp. 1860	0.23
Rhodothermus marinus DSM 4252	2.36
Riemerella anatipestifer ATCC 11845	3.61
Robiginitalea biformata HTCC2501	0.44
Salinibacter ruber M8	0.67
Sphaerobacter thermophilus DSM 20745	0.89
Sulfurimonas denitrificans DSM 1251	0.36
Sulfurovum sp. NBC37-1	0.16
Thermomicrobium roseum DSM 5159	1.47

Table B.3. Fraction of soil microbial community genomes encoding NosZ. The sequencing depth (i.e., average number of metagenomic reads spanning a nucleotide) for three housekeeping genes was used to estimate the fraction of the microbial community harboring *nosZ* genes at both agricultural sites.

	Havana sand		Urbana silt loam	
Gene	Seq. depth [bp]	*nosZ* fraction [%]	Seq. depth [bp]	*nosZ* fraction [%]
dnaK	4.27	14.6	9.77	15.1
recA	2.80	22.2	7.59	19.5
rpoB	5.48	11.4	9.40	15.7
nosZ	0.62		1.48	

Table B.4. Physicochemical properties of the agricultural soils studied.

Havana Sand depth [cm]	Total Organic Matter [%]	Available P [ppm-P]	K [ppm]	Mg [ppm]	Ca [ppm]	pH	NO_3^--N [ppm]	NH_4^+-N [ppm]
0 - 5	0.8	45	76	135	850	7.8	4	6
5 - 10	0.4	41	29	85	550	7.8	2	6
10 - 20	0.3	44	31	55	350	7.5	2	4
20 - 30	0.3	41	40	55	450	7.3	1	4

Urbana Silt loam depth [cm]	Total Organic Matter [%]	Available P [ppm-P]	K [ppm]	Mg [ppm]	Ca [ppm]	pH	NO_3^--N [ppm]	NH_4^+-N [ppm]
0 - 5	4	39	244	400	1950	5.9	4	9
5 - 10	4	23	128	395	2100	6.2	5	5
10 - 20	4	22	103	445	2400	6.3	5	5
20 - 30	3.8	11	61	415	2100	6.2	5	4

Table B.5 Agricultural soil metagenomes and assembly statistics. Assembly of short-reads was conducted by combining Velvet, SOAPdenovo and Newbler. Only contigs longer than 500 bp were considered for N50 calculations.

Soil Sample	Trimmed reads (sister reads) x 10^6	Coverage	Assembly (Kbp)			
			N50	Max	Fraction of assembled reads	Total
Havana sand	185.2	0.85	0.64	18	1.8%	38031
Urbana silt loam	247.02	0.40	0.77	16	5.3%	267501

APPENDIX C: SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Supporting Information C.1.

Methods

Functional annotation of short-reads using SEED in soil and fresh water metagenomes

SEED functional categories examined in detail for pathways of secondary metabolism included the terms “Iron acquisition and metabolism”, “Membrane transport”, “Metabolism of aromatic compounds”, “Motility and chemotaxis”, “Nitrogen metabolism”, “Phosphorus metabolism”, “Potassium metabolism”, “Secondary metabolism”, and “Sulfur metabolism”. Categories having above 0.01% relative abundance, on average, for top and deep soil layers in both sites were used for the determination of coefficient of variation between and within samples. The same annotation strategy was used for Lake Lanier metagenomes over the course of 1 year (1101B, 1104A, 1107A, and 1108A) and 2 years (1007B, 1008A, 1009A, 1010A, 1101B, 1104A, 1107A, and 1108A) (1).

Identification and analyses of 16S rRNA gene sequences

Short-read sequences encoding 16S rRNA gene fragments were extracted from each metagenome by using SortMeRNA (2) and their taxonomy was assigned using RDP classifier (cutoff 50)(3). Operational taxonomic units

(OTUs) were determined using a closed-reference OTU picking strategy as implemented in QIIME (4). Sequences were clustered into OTUs at 97% similarity using UCLUST (Edgar, 2010) and using references from SILVA database v111 (Quast *et al.*, 2013).

Identification of glycoside hydrolase genes

Glycoside hydrolase (GH) protein sequences in unassembled metagenomes were detected by querying the short-reads against the dbCAN database (5) using BLASTx (default settings and minimum 60% identity and 70% query coverage for a match). Bins harboring GH proteins were detected using BLASTp (default settings and minimum 60% identity and 70% query coverage for a match) against the previous database. In both cases, results were summarized based on the family classification from the CAZy database (6) and categories proposed previously (7).

Phylogenetic trees and placement of short-reads

Protein reference and assembled sequences were aligned using ClustalΩ (8) with default parameters. Resulting alignments were used to build phylogenetic trees in RAxML v8.0.19 (9). Identified short-reads encoding the protein of interest were extracted from soil metagenomes using ROCker (BLASTx) and their protein-coding sequences were predicted using FragGeneScan (10). The latter sequences were added to the corresponding protein alignment using MAFFT (“addfragments”) (11) and were placed in the corresponding phylogenetic tree using RAxML EPA (-f v option) (12).

Visualization and clade classification of reads placements

The visualization of the generated jplace files (13) was performed using the "JPlace.to_iToL.rb" script from the enveomics collection (14) and subsequently visualized on iTol (15). Quantification of the number of reads assigned to a specific clade (e.g., to distinguish between *nxrA* or *narG* reads) was done using the "JPlace.distances.rb" script, also available in the enveomics script collection.

To quantify *nirK* gene fragments assigned to specific clades we followed the clades previously proposed (16). The same process as described above for *nxrA/narG* was repeated except that all reads detected by ROCker models (I+II, III and *Thaumarchaeotea*) were used for classification. Clade IV (e.g., *Actinobacteria*) was intentionally omitted from this analysis due to the limited number of available genomes harboring *nirK*, which limited the development of a robust ROCker model. A similar clade specific approach was followed for the quantification of *nxrA* and *narG* gene fragments.

Results

Given the different amounts of OM observed between the two sites and soil layers, we hypothesized there would be site-specific microbial communities involved in the cycling and degradation of carbonaceous material. Specifically, we sought to find a link between the soil type and the dynamics of genes encoding enzymes (e.g., glycoside hydrolases) directly involved in the hydrolysis of glycosidic bonds in plant-derived carbon biomass. Despite the fact that genes

encoding glycoside hydrolases (GH) showed a slight increase (8%) at the end of the year in the top soil depth of Havana, stable GH gene abundances were observed within soil depths with respect to site (Appendix C, Figure C.9). For instance, GH genes encoding amylolytic enzymes showed high and stable abundance in both soils (up to 0.16% and 0.19% of total GH genes in Havana and Urbana, respectively), regardless of the differing soil texture and quantity of soil organic matter. Both sites showed significant higher relative abundance of GHs genes on the top compared to the deep soil layers (two tailed *t*-test, $p < 0.001$) (Appendix C, Figure C.9), and Urbana showed, on average, 20.4% higher relative abundance of GH genes compared to Havana.

Taxonomic compositions of agricultural soils

For Havana, *Proteobacteria* (~40%), *Acidobacteria* (~18%), and *Actinobacteria* (~17%) represented the most abundant phyla in both 0-5 and 20-30cm depths. *Bacteroidetes*, *Actinobacteria*, and *Firmicutes* were distinctive of the top soil metagenomes (*P*-value adjusted < 0.0001), whereas *Nitrospirae*, *Thaumarcheota*, and *Euryarchaeota* were characteristic of the deeper soil layer (*P*-value adjusted ≤ 0.001), in agreement with functional annotation results (Appendix C, Figure C.3b). At the order level, *Sphingomonadales*, *Sphingobacteriales*, *Actinomycetales*, and *Solirubrobacterales* were distinctive of the top layer, and *Nitrosopumilales*, *Neisseriales*, *Nitrospirales*, *Bacillales*, and *Rhodospirillales* were more abundant in the deeper metagenomes (*P*-value adjusted ≤ 0.0001). For Urbana, *Proteobacteria* (32%), *Actinobacteria* (22%) and *Acidobacteria* (~19%) represented the most abundant phyla in both depths

(Appendix C, Figure C.3b). *Bacteroidetes* and *Gemmatimonadetes* were more abundant in the top layer, whereas *Verrucomicrobia*, *Chloroflexi* and *Thaumarchaeota* were distinctive of the deep layer (*P*-value adjusted < 0.001). At the order level, *Flavobacteriales*, *Sphingomonadales*, *Caulobacterales*, *Xanthomonadales*, *Solirubrobacterales*, and *Burkholderiales* where characteristic of the top layer, whereas *Anaerolineales*, *Nitrospirales*, and *Nitrososphaerales* were distinctive of the lower layer (*P*-value adjusted < 0.05). Comparison of alpha diversity (Chao-Shen entropy index), based on the taxonomy at the phyla and order levels of the recovered 16S rRNA gene fragments, showed significant differences between the two soil layers in Urbana. For Havana, significant differences in alpha diversity were only detected at the phylum level between top and deep soils (Appendix C, Figure C.1b).

Using a closed reference OTU picking strategy, over 61% of the recovered 16S rRNA gene sequences in each site were clustered into an average of 3,482 and 2,170 OTUs (97% similarity) per sample in Havana and Urbana (defined at 97% 16S rRNA gene sequence identity). OTU projections per sample (Chao1 index) showed that Havana harbored more OTUs than Urbana soils (two-tailed *t*-test, $P < 0.01$). In addition, the latter estimates revealed that the detected OTUs in Havana ranged from 46% to 73% of the estimated total number of OTUs depending on the sample considered, whereas these values ranged from 49% to 82% in Urbana. Both sites shared 19.9% of the total detected OTUs (n=12,125) whereas 42.6% and 37.5% OTUs were specific to Havana and Urbana, respectively. A comparison of top vs. deep OTUs showed that in Havana,

statistically overrepresented OTUs (Log 2-fold >=2 and p-adjusted <0.01) in the top layer belonged to *Actinobacteria* (25.3%), *Alphaproteobacteria* (22.6%), and *Chloracidobacteria* (16.6%) whereas enriched OTUs in the deep layer belonged to *Gemmatimonadetes* (16%), *Nitrospirae* (10.2%), and *Thaumarchaeota* (10.2%). Similarly, overrepresented OTUs in the top layer of Urbana samples belonged to *Alphaproteobacteria* (46.5%), *Thermoleophilia* (14%) and *Actinobacteria* (11.6%) whereas enriched OTUs in the deep layer *Actinobacteria* (25.3%), *Alphaproteobacteria* (22.6%), and *Chloracidobacteria* (16.6%).

Denitrification genes

Hallmark denitrification genes showed stable abundances throughout the year but differences between soil layers and sites. For instance, in Havana, nitrate reductase (*narG*), nitrite reductases (*nirK* and *nirS*), and nitrous oxide reductase (*nosZ*) showed significant higher abundance in the deep compared to the top soil layer (Appendix C, Figure C.5). Even though both nitrite reductases were more abundant in the deeper soil layer, *nirK* was, on average, 9.5 and 6.1 times more abundant than *nirS* in the top and deep soil layers, respectively. On the other hand, opposite abundance patterns for denitrification genes were observed for Urbana. For instance, *narG*, *nirK*, *nirS*, and *norB* were statistically significantly more abundant in the surface soil layer compared to the deep soil layer (Appendix C, Figure C.5), probably as a result of the contrasting edaphic factors between sites. In addition, in both sites, *nrfA* showed the opposite abundance patterns compared to denitrification genes. Consistent with our previous reports from composite soil samples from the same agricultural soils (17), clade II, or

atypical *nosZ*, gene fragments showed higher abundance throughout the year in both sites. In Havana, clade II *nosZ* gene fragments were, on average, ~7 times more abundant than their clade I counterparts in both soil layers across the year. Interestingly, similar trends were observed in Urbana where atypical *nosZ* gene fragments were on average 9.7 and 15.9 times more abundant than their typical counterparts in the top and deep soil layers throughout the year, respectively.

Recovered populations from metagenomes

The assembly and binning resulted in 69 population bin genomes in total from both sites, having over 50% completion and less than 20% of contamination based on the presence of 104 and 26 single-copy bacterial and archaeal genes, respectively. These genes might not always be present in all microbial lineages, therefore, gene content and completeness values were likely underestimated. The use of relatively low stringency criteria was due the low fraction of assembled metagenomic reads. Even at this level of stringency, only 69 bins, representing ~30% of the total bins obtained, were selected. In fact, the remaining bins were even more incomplete or contaminated despite efforts to refine binning by performing a second round of assembly (see Material and Methods for details). The majority of the 69 bins were obtained from the 20-30 cm layer (Havana =45/47, Urbana=18/22) presumably due to the lower sequence diversity and higher average coverage as determined by Nonpareil. Genome sizes ranged from 1.1 to 6.7 Mbp, and G+C% content varied from 35 to 72% (Appendix C, Table C.6). Inferred taxonomy revealed that most bins represented members of *Proteobacteria*, *Acidobacteria*, and *Actinobacteria* in both soils

whereas *Verrucomicrobia* and *Gemmatimonadetes* were characteristic of Urbana and Havana, respectively. As expected, genomic comparisons based on average amino acid (AAI) values (18) revealed that most of the obtained bins likely represented novel organisms when compared to the NCBI prokaryotic genome database (Appendix C, Table C.6). For Havana, only 4.3% of the bins had AAI values greater than ~65% (i.e., shared genus) (19) compared to their close relatives. A similar trend was observed for Urbana bins where none of the bins likely corresponded to known genera. However, closely related bins, most likely representing member of the same genus (i.e., sharing AAI >65%), were detected in both sites. For instance, in Havana, *Nitrospira* bins HD017 and HD021 shared 81.69% AAI (SD: 15.43%, from 2288 proteins); *Gemmatimonadetes* bins HD002 and HD027 shared 77.47% AAI (SD: 15.80%, from 2429 proteins). In Urbana, *Verrucomicrobia* bins UD002 and UD007 shared 82.65% AAI (SD: 16.82%, from 1713 proteins). Several bins were specific to each site but shared relatively high AAI values such as the *Thaumarchaeota* bins HD032 and UD001 which shared 76.79% AAI (SD: 14.46%, from 1560 proteins).

Diversity of population bin genomes involved in carbon cycling

Differences in the number of genes encoding key polysaccharide degradation enzymes (i.e., glycoside hydrolase enzymes) were observed between the population bins. For instance, bins from Urbana encode significantly more glycoside hydrolases (GH) compared to Havana bins (unpaired *t*-test, *P*-value < 0.05, see also Appendix C, Table C.7). In addition, bins from Urbana showed almost double the number of cellulase genes encoding oligosaccharide-

degrading enzymes and amylolitic enzymes compared to bins from Havana. Genes encoding beta-glucosidase enzymes GH3 (n=93) and the amylolytic enzymes GH13 (n=320) and GH15 (n=78), were among the most commonly detected glycoside hydrolases in recovered bins. These results were consistent with the results obtained from recovered short-reads and, in general, with a higher soil organic matter content in the Urbana (silty loam) soil vs. its Havana (sandy) counterpart. The bins UD035 (*Actinobacteria*), UD029 (*Firmicutes*), and UT009 (*Acidobacteria*) from Urbana had the highest number of GH genes (n=67, 41 and 49 GH genes) and mostly corresponded to oligosaccharide-degrading and amylolytic enzymes. In Havana, bins HD112 and HD089 (*Acidobacteria*) and bin HD116 (*Bacteroidetes*) had the highest number of HG genes also corresponding to cellulases, oligosaccharide-degrading and amylolytic enzymes.

Discussion

Unexpected genetic diversity in agricultural soils

The majority of the bins were predicted to belong to novel species, if not genera, reflecting the low representation of soil-dwelling microorganisms in current genomic databases. For instance, highly abundant archaeal and bacterial nitrifier (discussed above) and *Verrucomicrobia* populations obtained from Urbana (e.g., bin UD002) only shared ~46% AAI to the closest referential genome. Microbial communities belonging to the latter group are underrepresented in genomic databases and have been predicted to inhabit soils with high organic matter content such as those found in Urbana (20). It is

important to note that while the bin genomes were searched against the NCBI prokaryotic genome database (as implemented in MiGA) for close relatives, more recently sequenced genomes, which are not yet part of NCBI, would have been missed. For instance, bin UD053 shared 61% AAI with recently described and novel phylum *Candidatus* Rokubacteri*a* (21). Further, abundant populations detected based on 16S rRNA gene fragments recovered in the metagenomes were not well represented in the recovered bin genomes, such as *Gemmatimonadetes* in Urbana. Apparently, the latter genomes were not well binned, presumably due to high intra-population sequence diversity. Altogether, the genomic bins reported here represent mostly novel and deep-branching taxa and offer a genomic reference for future studies targeting abundant natural microbial communities found in agricultural soils.

Recent findings have revealed that PCR-based surveys targeting N-cycle genes have overlooked a vast amount of natural diversity related to nitrification (22-25) and denitrification genes such as *nirK* (16)*, nosZ* (17), and *nrfA* (26). These findings suggest that many of the previously unaccounted gene diversity might play an important role in key biogeochemical cycles. Our results show that the use of metagenomic approaches in combination with reliable detection tools (e.g., ROCker) can circumvent these limitations in samples of high sequence complexity. For instance, abundant *nirK* genes found in the soil samples were assigned to *Thaumarchaeota*, which has been inadvertently excluded in previous PCR-based gene surveys. Interestingly, the changes in relative abundance for *Thaumarchaeota nirK* gene fragments are consistent with recent findings that

have proposed an alternative role for this archaeal NirK activity as part of the ammonia oxidation to nitrite mechanism in *Thaumarchaeota* (27).

Genes and microbial populations involved in biomass degradation

We explored the impact of the microbial communities in the breakdown and recycling of plant biomass in soils, by surveying genes associated with biomass and polysaccharide degradation. For instance, glycoside hydrolases (GH) are a widespread group of enzymes that catalyze the hydrolysis of the glycoside bond and are key for degradation of labile, e.g., starch and polysaccharides, as well as recalcitrant, e.g., lignocellulose and other complex organic carbon compounds (6). The two agricultural sites share a similar history of cropping where biomass derived from either corn or soybean represents a constant input of C at the end of the growing season and this was reflected by stable abundances in all GH categories studied. Even though a higher influence of plants (e.g., root-exudates) during crop-growing periods was expected (e.g., June and September), our core collecting regime was directed to sample in between plant rows, and thus, likely missed microorganisms in close proximity to roots. Overall, Urbana (silt-loam soil) showed a higher relative abundance of GH genes at both gene and genomic population levels compared to Havana, likely explained by the intrinsic characteristics of the soils. For instance, the differences in sorption and binding capacities particular to each soil type resulted in a higher OM availability in Urbana compared to Havana, which likely accounted for the differences in GH genes between the two sites. Further, previous reports have recognized that genes encoding GHs belonging to the family GH13 are among

the most widespread and abundant amylolytic enzymes found in microbial genomes (28) and soils (29), consistent with the findings based on the recovered bin genome sequences reported here. Other abundant GHs in the recovered bins belonged to the glucoamylase GH15 family, which in combination with debranching enzymes from GH13 have been proposed as part of the main enzymes for degradation of polysaccharides in bacteria (28). Therefore, in addition to playing a role in the cycling of N in soils, bins encoding GH might also participate in maintaining and recycling labile carbon in the explored agricultural soils.

References

1. **Oh S**, **Caro-Quintero A**, **Tsementzi D**, **DeLeon-Rodriguez N**, **Luo C**, **Poretsky R**, **Konstantinidis KT**. 2011. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. Appl Environ Microbiol **77**:6000–6011.

2. **Kopylova E**, **Noé L**, **Touzet H**. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics **28**:3211–3217.

3. **Wang Q**, **Garrity GM**, **Tiedje JM**, **Cole JR**. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol **73**:5261–5267.

4. **Caporaso JG**, **Kuczynski J**, **Stombaugh J**, **Bittinger K**, **Bushman FD**, **Costello EK**, **Fierer N**, **Peña AG**, **Goodrich JK**, **Gordon JI**, **Huttley GA**, **Kelley ST**, **Knights D**, **Koenig JE**, **Ley RE**, **Lozupone CA**, **McDonald D**, **Muegge BD**, **Pirrung M**, **Reeder J**, **Sevinsky JR**, **Turnbaugh PJ**, **Walters WA**, **Widmann J**, **Yatsunenko T**, **Zaneveld J**, **Knight R**. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods **7**:335–336.

5. **Xu Y**. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res **40**:W445–51.

6. **Lombard V**, **Golaconda Ramulu H**, **Drula E**, **Coutinho PM**, **Henrissat B**. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res **42**:D490–5.

7. **Allgaier M**, **Reddy A**, **Park JI**, **Ivanova N**, **D'haeseleer P**, **Lowry S**, **Sapra R**, **Hazen TC**, **Simmons BA**, **VanderGheynst JS**, **Hugenholtz P**. 2010. Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community. PLoS ONE **5**:e8812.

8. **Sievers F**, **Wilm A**, **Dineen D**, **Gibson TJ**, **Karplus K**, **Li W**, **Lopez R**, **McWilliam H**, **Remmert M**, **Söding J**, **Thompson JD**, **Higgins DG**. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol **7**:539–539.

9. **Stamatakis A**. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688–2690.

10. **Rho M**, **Tang H**, **Ye Y**. 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res **38**:e191.

11. **Katoh K**, **Standley DM**. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30**:772–780.

12. **Berger SA**, **Krompass D**, **Stamatakis A**. 2011. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. Syst Biol **60**:291–302.

13. **Matsen FA**, **Hoffman NG**, **Gallagher A**, **Stamatakis A**. 2012. A format for phylogenetic placements. PLoS ONE **7**:e31009.

14. **Rodriguez-R LM**, **Konstantinidis KT**. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints **4**:e1900v1.

15. **Letunic I**, **Bork P**. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res **39**:gkr201–W478.

16. **Wei W**, **Isobe K**, **Nishizawa T**, **Zhu L**, **Shiratori Y**, **Ohte N**, **Koba K**, **Otsuka S**, **Senoo K**. 2015. Higher diversity and abundance of denitrifying microorganisms in environments than considered previously. The ISME Journal **9**:1–12.

17. **Orellana LH**, **Rodriguez-R LM**, **Higgins S**, **Chee-Sanford JC**, **Sanford RA**, **Ritalahti KM**, **Löffler FE**, **Konstantinidis KT**. 2014. Detecting nitrous oxide reductase (NosZ) genes in soil metagenomes: method development

and implications for the nitrogen cycle. mBio **5**:e01193–14.

18. **Konstantinidis KT**. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci USA **102**:2567–2572.

19. **Konstantinidis KT**, **Tiedje JM**. 2007. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Current Opinion in Microbiology **10**:504–509.

20. **Brewer TE**, **Handley KM**, **Carini P**, **Gilbert JA**, **Fierer N**. 2016. Genome reduction in an abundant and ubiquitous soil bacterium 'Candidatus Udaeobacter copiosus'. Nat Microbiol **2**:16198.

21. **Hug LA**, **Thomas BC**, **Sharon I**, **Brown CT**, **Sharma R**, **Hettich RL**, **Wilkins MJ**, **Williams KH**, **Singh A**, **Banfield JF**. 2015. Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. Environmental Microbiology **18**:159–173.

22. **Daims H**, **Lebedeva EV**, **Pjevac P**, **Han P**, **Herbold C**, **Albertsen M**, **Jehmlich N**, **Palatinszky M**, **Vierheilig J**, **Bulaev A**, **Kirkegaard RH**, **Bergen von M**, **Rattei T**, **Bendinger B**, **Nielsen PH**, **Wagner M**. 2015. Complete nitrification by Nitrospira bacteria. Nature **528**:504–509.

23. **van Kessel MAHJ**, **Speth DR**, **Albertsen M**, **Nielsen PH**, **Op den Camp HJM**, **Kartal B**, **Jetten MSM**, **Lücker S**. 2015. Complete nitrification by a single microorganism. Nature **528**:555–559.

24. **Palomo A**, **Fowler SJ**, **Gülay A**, **Rasmussen S**, **Sicheritz-Ponten T**, **Smets BF**. 2016. Metagenomic analysis of rapid gravity sand filter microbial communities suggests novel physiology of Nitrospira spp. The ISME Journal **10**:2569–2581.

25. **Pinto AJ**, **Marcus DN**, **Ijaz UZ**, **Bautista-de Lose Santos QM**, **Dick GJ**, **Raskin L**. 2016. Metagenomic Evidence for the Presence of Comammox Nitrospira-Like Bacteria in a Drinking Water System. mSphere **1**:e00054–15.

26. **Nelson MB**, **Martiny AC**, **Martiny JBH**. 2016. Global biogeography of microbial nitrogen-cycling traits in soil. Proc Natl Acad Sci USA **113**:8033–8040.

27. **Kozlowski JA**, **Stieglmeier M**, **Schleper C**, **Klotz MG**, **Stein LY**. 2016. Pathways and key intermediates required for obligate aerobic ammonia-dependent chemolithotrophy in bacteria and Thaumarchaeota. The ISME Journal 1–10.

28. **Henrissat B**, **Deleury E**, **Coutinho PM**. 2002. Glycogen metabolism loss:

a common marker of parasitic behaviour in bacteria? Trends in Genetics **18**:437–440.

29. **Howe A**, **Yang F**, **Williams RJ**, **Meyer F**, **Hofmockel KS**. 2016. Identification of the Core Set of Carbon-Associated Genes in a Bioenergy Grassland Soil. PLoS ONE **11**:e0166578.

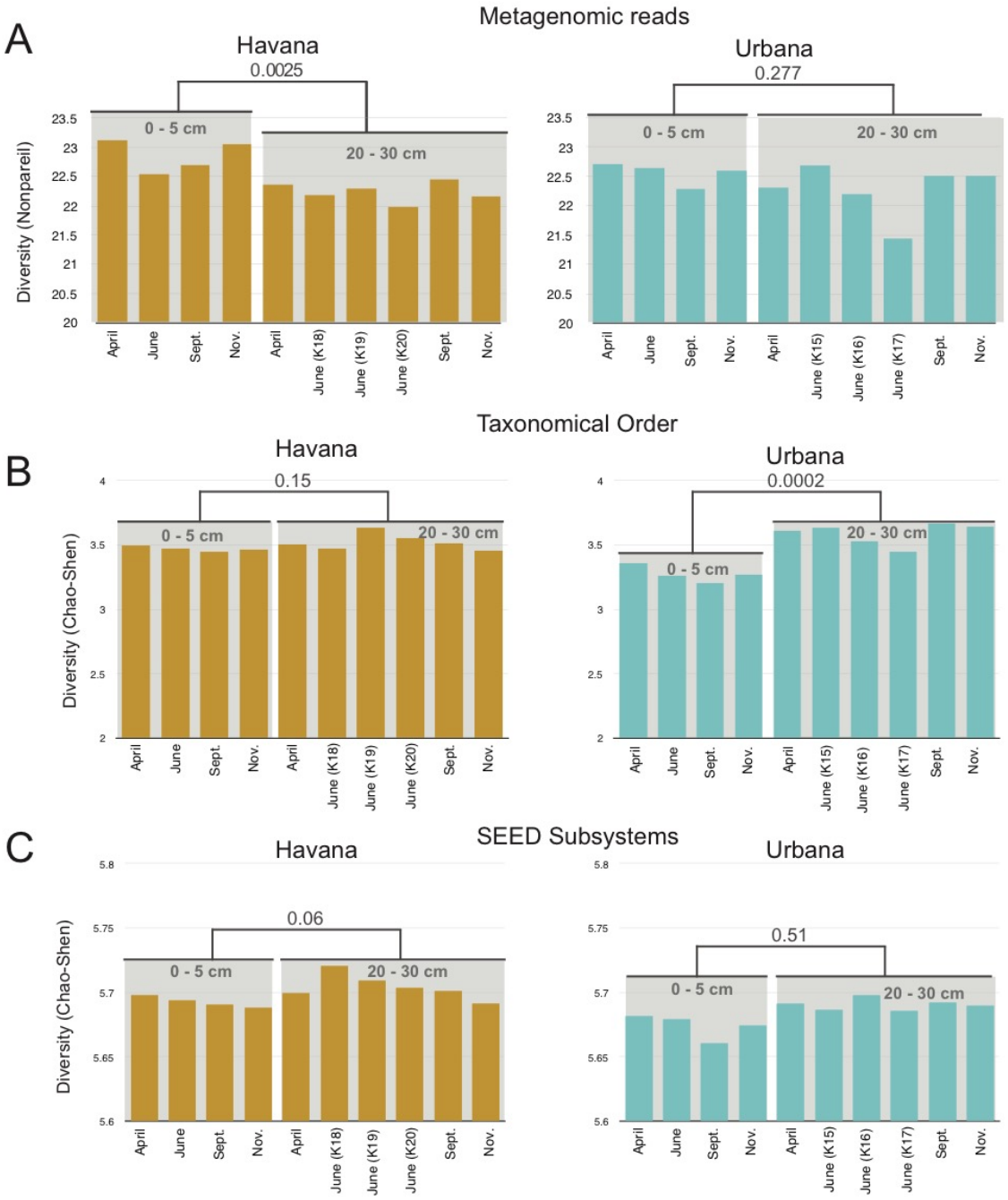


Figure C.1. Alpha diversity values determined for metagenomic samples. **A.** Diversity of metagenomic reads as determined by Nonpareil. The Chao-Shen entropy values for **B.** the order level of taxonomy and **C.** functional annotations (SEED subsystems).

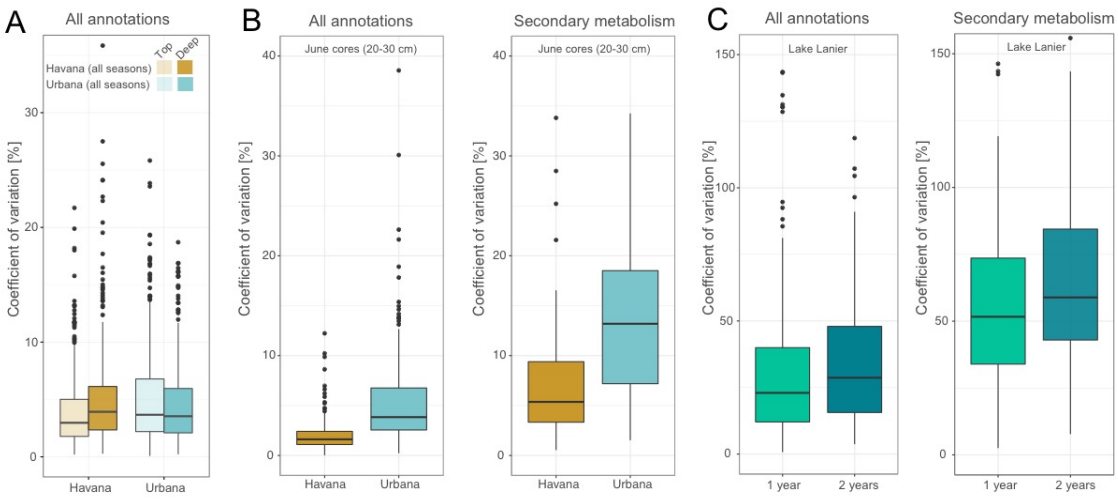


Figure C.2. A. Distributions of coefficients of variation for all SEED subsystems detected in soil metagenomes for all seasons. Panel B summarizes the distributions of coefficient of variation for all SEED subsystems (left) and the subset devoted to secondary metabolism (right) for the three cores obtained for the 20-30 cm soil samples during June in Havana and Urbana. **C.** Distributions of coefficients of variation for all SEED subsystems (left) and a subset consisting of secondary metabolism annotations (right panel) in Lake Lanier metagenomes.

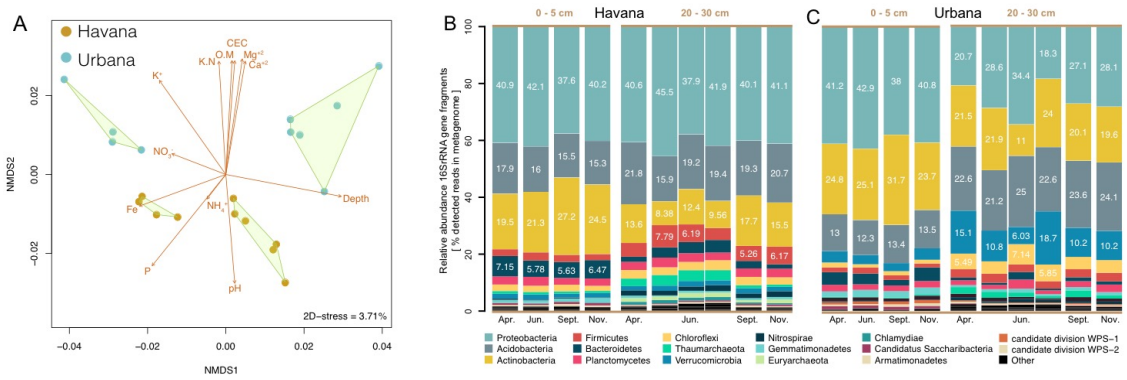


Figure C.3. Functional clustering and taxonomy for sandy (Havana) and silt-loam (Urbana) soils. A. Non-metric multidimensional scaling analysis based on SEED subsystems annotation of short-reads of the metagenomic samples showed independent clustering by site and depth. The length of the arrow is proportional to the correlation between measured metadata and determined ordination values. **B** Summary of the taxonomic affiliation (figure key) of the recovered 16S rRNA gene fragments obtained from soil metagenomes.

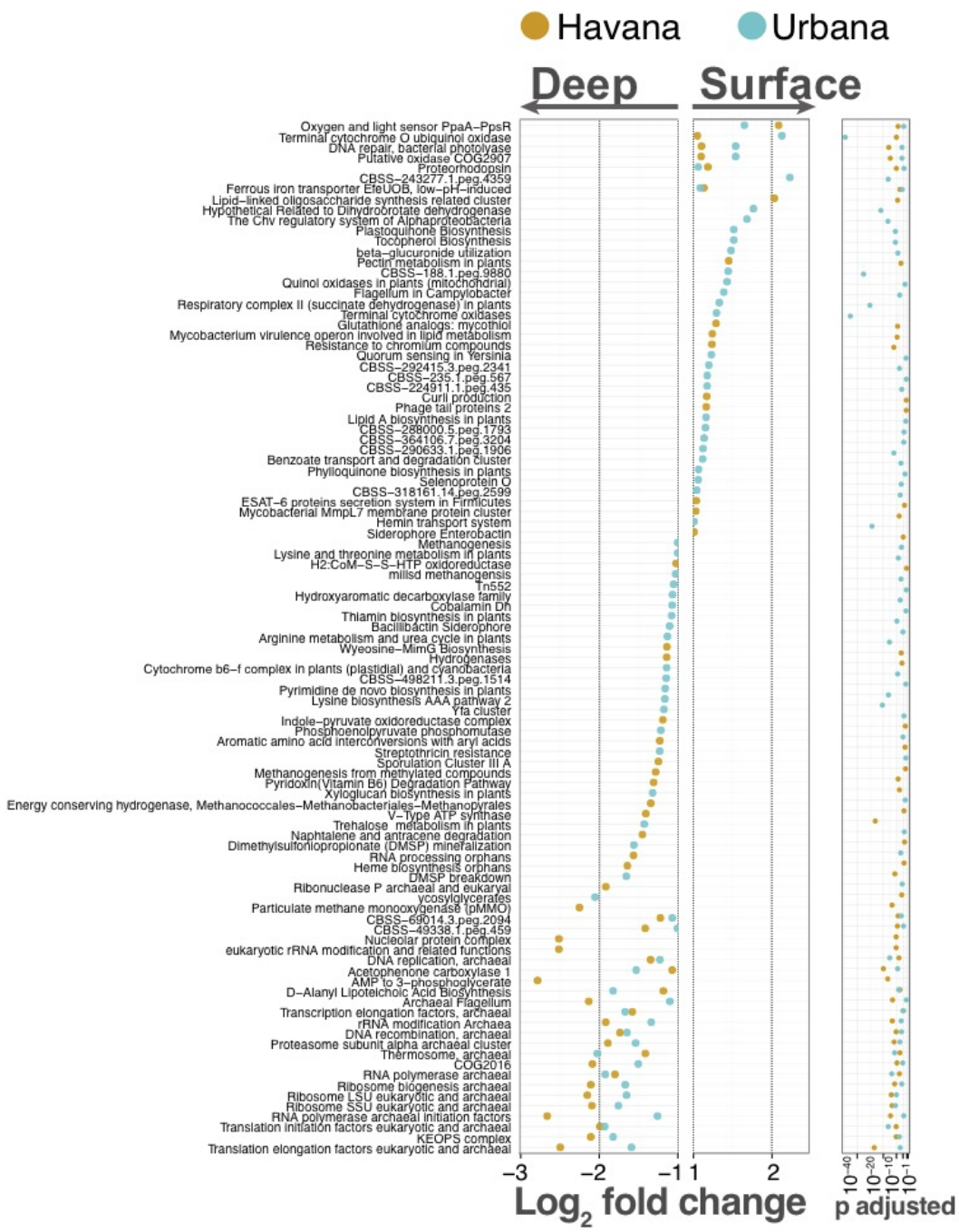


Figure C.4. Differential abundance of SEED subsystems between top (0-5cm) and deep (20-30 cm) soil layers. Predicted-protein sequences from short-reads were annotated using UniProt and subsequently classified into functional categories using SEED subsystems. Significant differences in abundance of

SEED subsystems between top and deep layers were identified using a negative binomial test as implemented in DESeq2. Selected SEED subsystems showing log_2-fold change >=1 or <=-1 and adjusted *P*-values < 0.01 are shown.

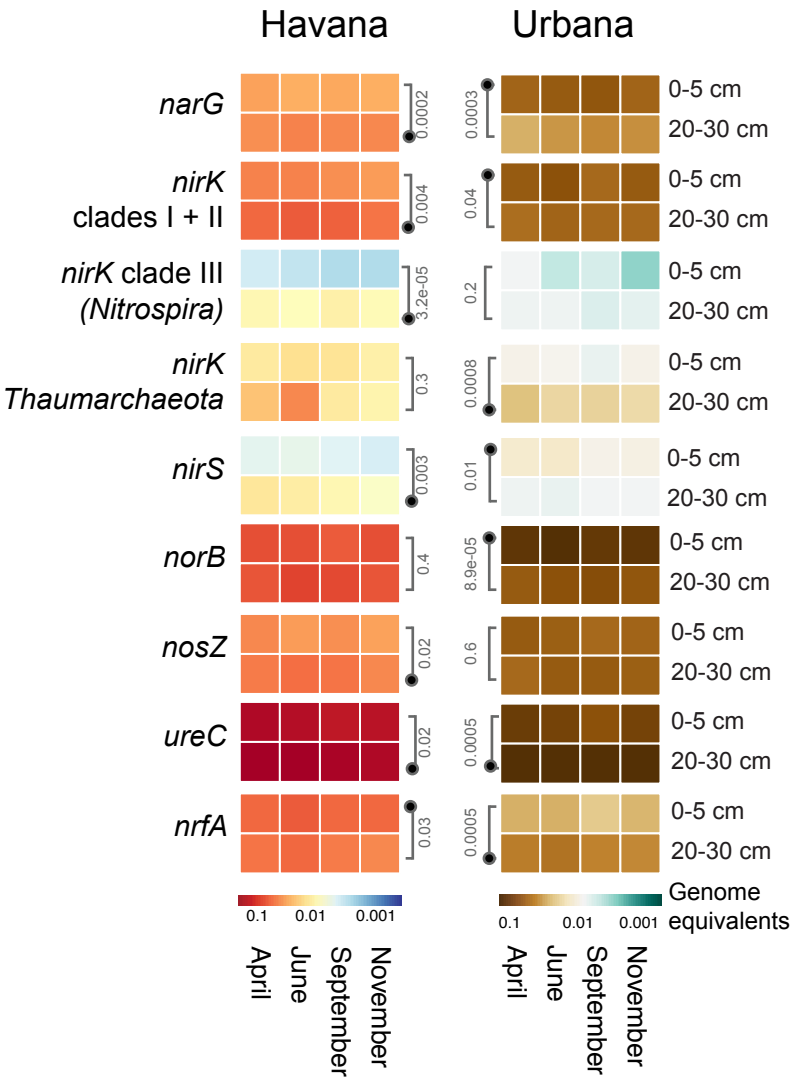


Figure C.5. Abundance of N-cycle genes in sandy (Havana) and silt-loam (Urbana) soils. Heatmaps show calculated relative abundance for N-cycle genes as genome equivalents for Havana (left panel) and Urbana (right panel). Manually-curated databases for each N gene were searched against soil metagenomes using BLASTx and outputs were filtered using ROCker models for each gene (see Methods for more details). Values for the 20-30 cm layer in June represent the average of the three soil cores.

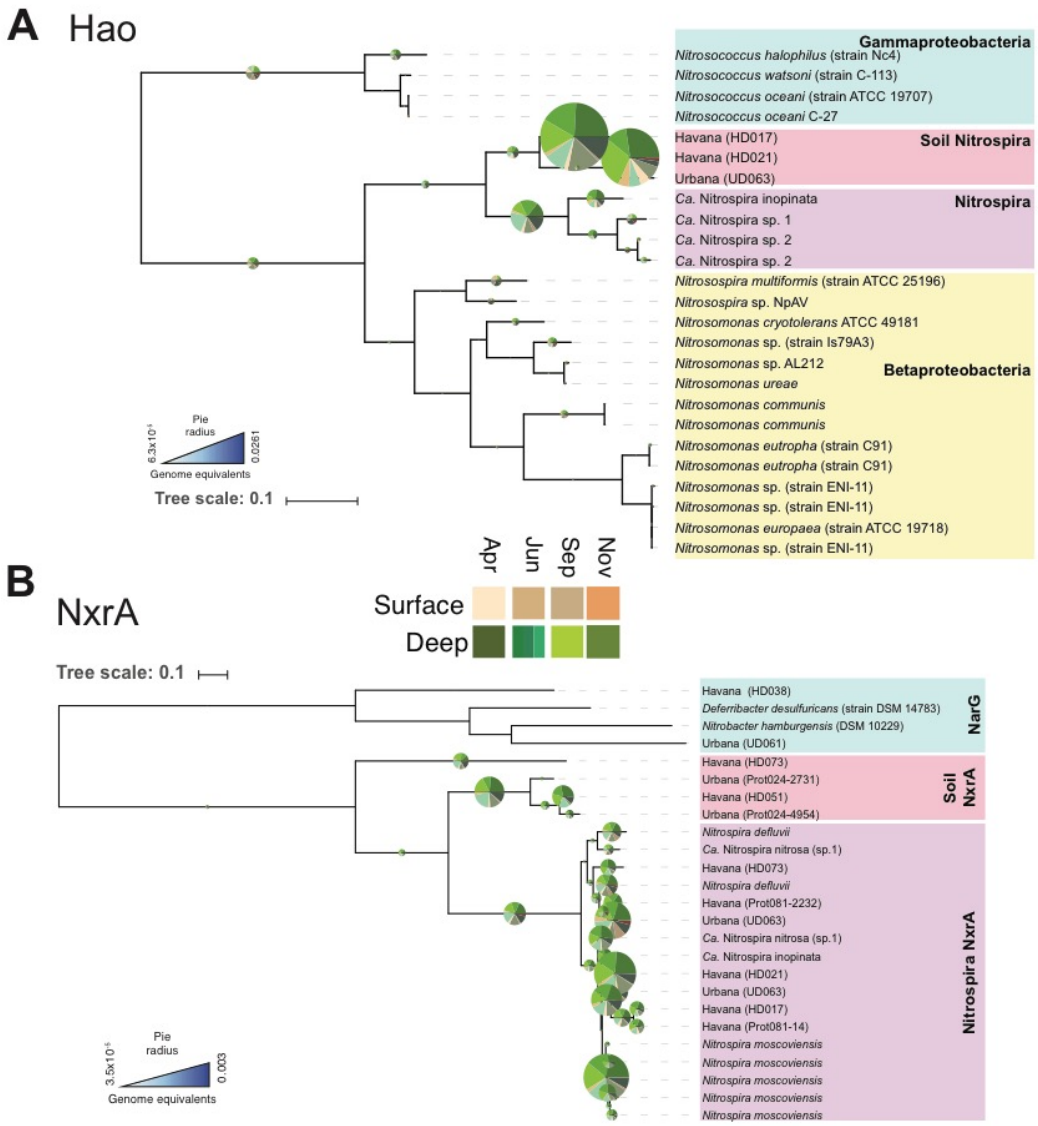
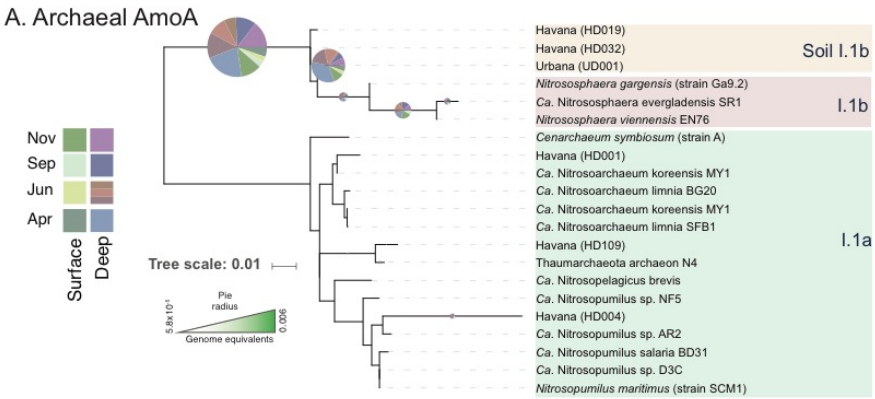
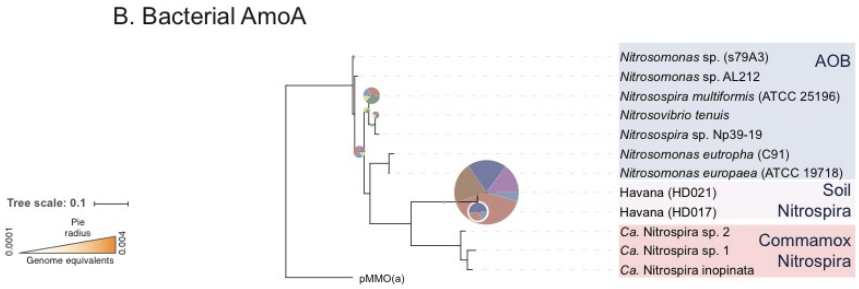
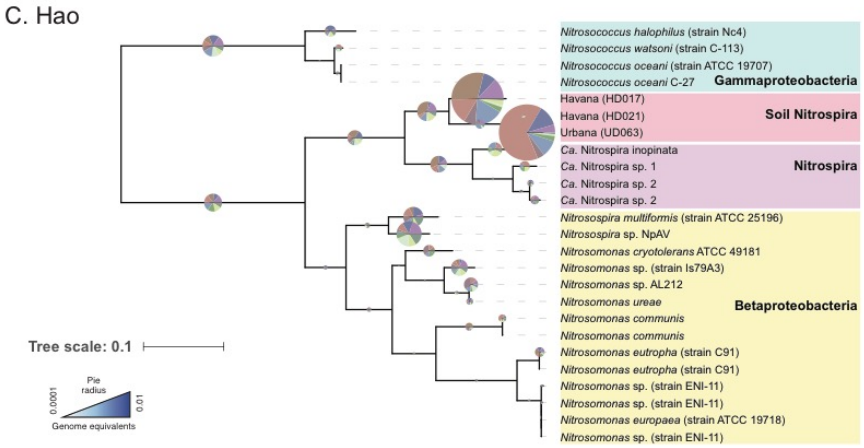


Figure C.6. Abundance and diversity for *hao* and *nxrA* in Havana. Phylogenetic reconstruction of Hao (**A**) and NxrA (**B**) protein sequences including assembled sequences from soil metagenomes. For reconstructed sequences, names in parentheses indicate corresponding metagenomic bins. The pie charts represent the placing of Havana metagenomic reads for archaeal and bacterial *amoA* genes using RAxML EPA. Pie chart radii represent the read abundance for

each node (calculated as genome equivalents) and the colors of the slices represent the depth and month for the origin of the metagenomic reads.


A. Archaeal AmoA
Nov
Sep
Jun
Apr
Surface
Deep
Tree scale: 0.01
Pie radius
5.8x10⁻¹
0.005
Genome equivalents
Havana (HD019)
Havana (HD032)
Urbana (UD001)
Soil I.1b
Nitrososphaera gargensis (strain Ga9.2)
Ca. Nitrososphaera evergladensis SR1
Nitrososphaera viennensis EN76
I.1b
Cenarchaeum symbiosum (strain A)
Havana (HD001)
Ca. Nitrosoarchaeum koreensis MY1
Ca. Nitrosoarchaeum limnia BG20
Ca. Nitrosoarchaeum koreensis MY1
Ca. Nitrosoarchaeum limnia SFB1
Havana (HD109)
Thaumarchaeota archaeon N4
Ca. Nitrosopelagicus brevis
Ca. Nitrosopumilus sp. NF5
Havana (HD004)
Ca. Nitrosopumilus sp. AR2
Ca. Nitrosopumilus salaria BD31
Ca. Nitrosopumilus sp. D3C
Nitrosopumilus maritimus (strain SCM1)
I.1a


B. Bacterial AmoA
Tree scale: 0.1
Pie radius
0.0001
0.004
Genome equivalents
pMMO(a)
Nitrosomonas sp. (s79A3)
Nitrosomonas sp. AL212
Nitrosospira multiformis (ATCC 25196)
Nitrosovibrio tenuis
Nitrosospira sp. Np39-19
Nitrosomonas eutropha (C91)
Nitrosomonas europaea (ATCC 19718)
AOB
Havana (HD021)
Havana (HD017)
Soil Nitrospira
Ca. Nitrospira sp. 2
Ca. Nitrospira sp. 1
Ca. Nitrospira inopinata
Commamox Nitrospira


C. Hao
Tree scale: 0.1
Pie radius
0.0001
0.01
Genome equivalents
Nitrosococcus halophilus (strain Nc4)
Nitrosococcus watsoni (strain C-113)
Nitrosococcus oceani (strain ATCC 19707)
Nitrosococcus oceani C-27
Gammaproteobacteria
Havana (HD017)
Havana (HD021)
Urbana (UD063)
Soil Nitrospira
Ca. Nitrospira inopinata
Ca. Nitrospira sp. 1
Ca. Nitrospira sp. 2
Ca. Nitrospira sp. 2
Nitrospira
Nitrosospira multiformis (strain ATCC 25196)
Nitrosospira sp. NpAV
Nitrosomonas cryotolerans ATCC 49181
Nitrosomonas sp. (strain Is79A3)
Nitrosomonas sp. AL212
Nitrosomonas ureae
Nitrosomonas communis
Nitrosomonas communis
Nitrosomonas eutropha (strain C91)
Nitrosomonas eutropha (strain C91)
Nitrosomonas sp. (strain ENI-11)
Nitrosomonas sp. (strain ENI-11)
Nitrosomonas europaea (strain ATCC 19718)
Nitrosomonas sp. (strain ENI-11)
Betaproteobacteria

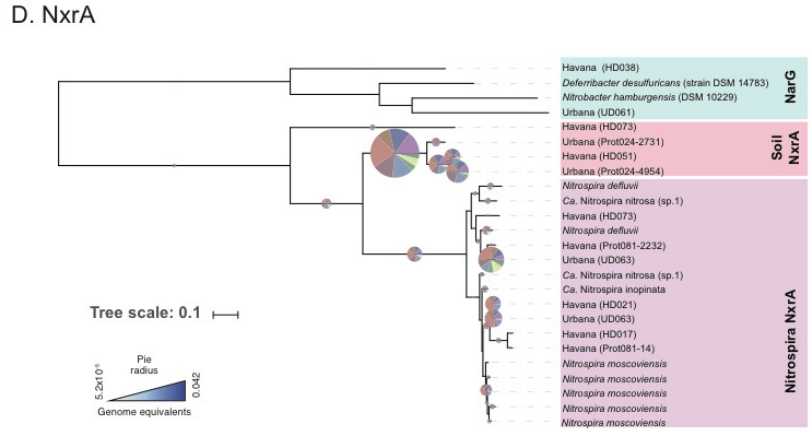

D. NxrA
Tree scale: 0.1
Pie radius
5.2x10⁻¹
0.042
Genome equivalents
Havana (HD038)
Deferribacter desulfuricans (strain DSM 14783)
Nitrobacter hamburgensis (DSM 10229)
Urbana (UD061)
NarG
Havana (HD073)
Urbana (Prot024-2731)
Havana (HD051)
Urbana (Prot024-4954)
Soil NxrA
Nitrospira defluvii
Ca. Nitrospira nitrosa (sp.1)
Havana (HD073)
Nitrospira defluvii
Havana (Prot081-2232)
Urbana (UD063)
Ca. Nitrospira nitrosa (sp.1)
Ca. Nitrospira inopinata
Havana (HD021)
Urbana (UD063)
Havana (HD017)
Havana (Prot081-14)
Nitrospira moscoviensis
Nitrospira moscoviensis
Nitrospira moscoviensis
Nitrospira moscoviensis
Nitrospira moscoviensis
Nitrospira NxrA

Figure C.7. Abundance and diversity for archaeal and bacterial *amoA, hao,* and *nxrA* in Urbana. Phylogenetic reconstruction of archaeal (**A**) and bacterial (**B**) AmoA, Hao (**C**) and NxrA (**D**) protein sequences including assembled sequences from soil metagenomes. For reconstructed sequences, names inside parentheses indicate corresponding metagenomic bins. The pie charts represent the placement of Havana metagenomic reads for archaeal and bacterial *amoA* genes using RAxML EPA. Pie chart radii represent the read abundance for each node (calculated as genome equivalents) and the colors of the slices represent the depth and month for the origin of the metagenomic reads.

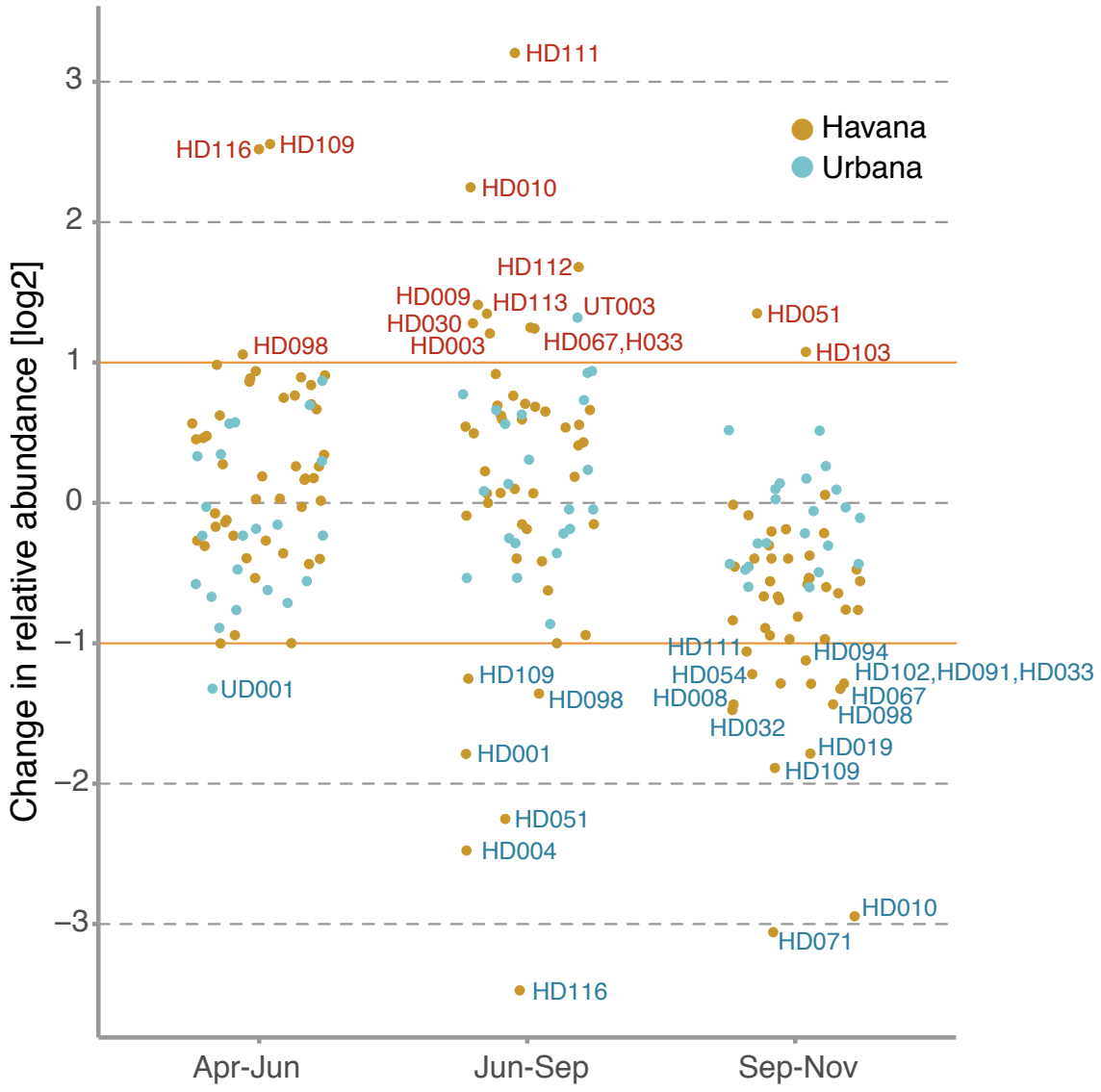


Figure C.8. Changes in abundance of metagenomic populations. Log_2 fold changes in abundance (y-axis) between months (x-axis) were calculated using individual bin abundances.

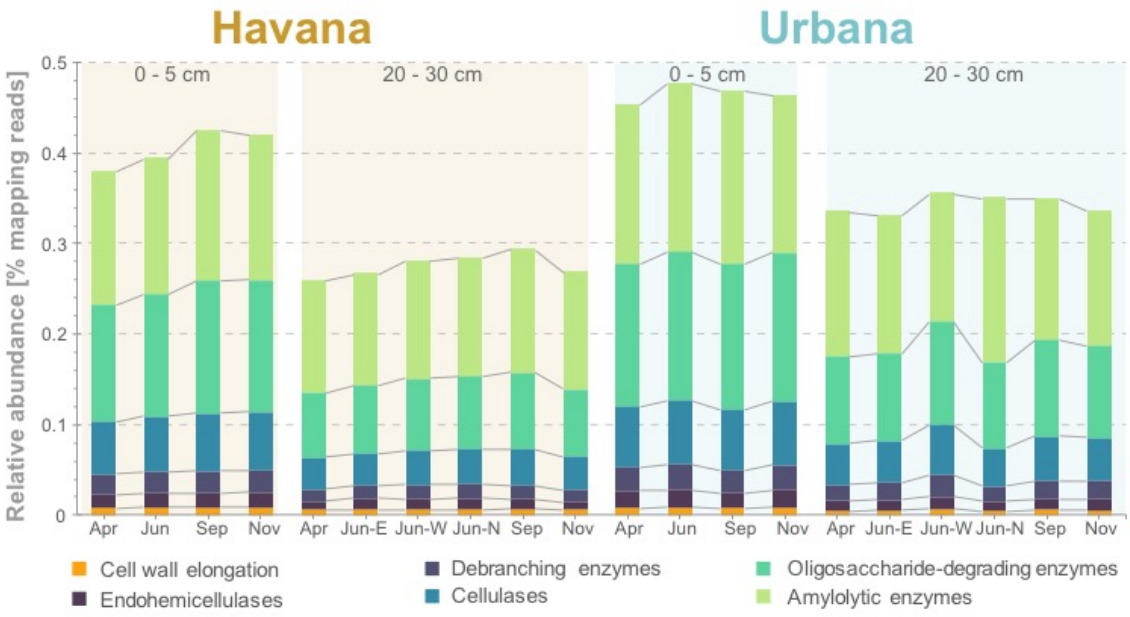


Figure C.9. Relative abundances of categories for glycoside hydrolases in both agricultural soils. Glycoside hydrolases (GH) gene fragments were detected in each metagenome and individual GH abundances were summarized in six functional categories.

Table C.1. Soil metadata for Havana and Urbana samples.

Site	ID	Depth	Sampling date during 2012	Soil metadata												
				pH	Total organic matter	Available P	K	Mg	Ca	NO_3^--N	NH_4^+-N	Total Kjeldahl N	Extractable Fe	CEC	Temp	Moisture
		[cm]			[%]	[ppm-P]	[ppm]	[ppm]	[ppm]	[ppm]	[ppm]	[%]	[ppm]	[meq/100g]	[°C]	[%]
Havana (sandy soil)	K10	0 - 5	Apr 4	7.7	0.7	49	59	118	729	4	5	0.039	160	4.8	17.9	4.63
	K14		Jun 6	7.7	1.3	55	91	126	838	74	139	0.083	138	5.5	32.6	4.99
	K6		Sep 5	7.3	1.2	53	68	144	851	6	2	0.064	150	5.6	23.1	6.33
	K2		Nov 6	7.6	0.9	54	78	139	816	8	5	0.048	142	5.4	8.4	4.65
	K12	20 - 30	Apr 4	7.4	0.4	43	41	60	443	1	2	0.02	132	2.8	17.4	4.93
	K18 (E)		Jun 6	7.35	0.6	50	38	56	572	1	4	0.022	163	3.4	22.8	6.74
	K19 (M)		Jun 6	7.3	0.6	50	38	56	572	1	4	0.022	163	3.4	22.8	3.53
	K20 (W)		Jun 6	7.48	0.6	50	38	56	572	1	4	0.022	163	3.4	22.8	5.49
	K8		Sep 5	7	0.3	48	49	70	489	1	1	0.021	160	3.2	23.6	4.85
	K4		Nov 6	7.4	0.4	58	43	77	471	1	3	0.027	162	3.1	10	4.55
Urbana (silt-loam soil)	K9	0 - 5	Apr 2	5.9	3.7	46	179	355	1,800	12	4	0.158	172	18.6	22.7	19.31
	K13		Jun 4	5.3	3.5	45	202	369	1,998	26	4	0.16	161	19.6	20.4	18.65
	K5		Aug 29	5.6	4.2	54	234	425	2,412	29	3	0.167	194	21	21.3	19.33
	K1		Nov 8	6	3.7	46	187	373	2,051	6	4	0.152	182	17.4	9.1	21.82
	K11	20 - 30	Apr 2	6.2	4.1	21	122	558	3,135	4	5	0.15	138	24.2	18.2	21.6
	K15 (M)		Jun 4	5.92	3.8	18	59	456	2,550	7	4	0.14	113	19.1	19.7	20.33
	K16 (S)		Jun 4	6.92	3.8	18	59	456	2,550	7	4	0.14	113	19.1	19.7	19.71
	K17 (N)		Jun 4	6.2	3.8	18	59	456	2,550	7	4	0.14	113	19.1	19.7	22.69
	K7		Aug 29	6.1	4.1	25	102	498	2,913	4	3	0.155	138	22.6	21.8	18.78
	K3		Nov 8	6.2	3.7	20	79	449	2,669	7	4	0.127	126	20.9	7.8	20.93

Table C.2. Agricultural management for Havana and Urbana sites during 2012.

Site	Sampling date during 2012	Crop information	Tillage & N-fertilizer input	Notes
Havana	Apr 4	Pre-tillage, pre-fertilizer, pre-planting at time of sampling (winter fallow)	Pre-Tillage, Pre-fertilizer	
	Jun 6	Corn planted May 12	Spring tillage, UAN28 applied late April (180 lb N/acre)	Herbicide applied June 15
	Sep 5	Full canopy corn; beginning senesce		
	Nov 6	Post-soybean harvest by time of sampling; harvested few days prior		
Urbana	Apr 2	Pre-planting at time of sampling (winter fallow)	Pre-Tillage	
	Jun 4	Pre-planting at time of sampling	Spring tillage No UAN28 application this crop year	Soybean planted Jun 6, 2012, glyphosate late June
	Aug 29	Full canopy soybean		Full growing season rain-fed only
	Nov 8	Post-harvest; Soybean harvested Nov 1	No Fall tillage yet	

Table C.3. Metagenomic sequences and Nonpareil estimations for Havana and Urbana sites.

Site	ID	Depth	Month	Sequences		Coverage
				Trimmed Reads*	Trimmed reads length (avg)	
Havana	K2	0 - 5 cm	November	27,808,182	123.8	0.117
	K4	20 - 30 cm	November	24,373,825	123.7	0.192
	K6	0 - 5 cm	September	32,787,009	124.3	0.116
	K8	20 - 30 cm	September	33,047,556	124.7	0.178
	K10	0 - 5 cm	April	38,337,187	124.5	0.105
	K12	20 - 30 cm	April	29,017,415	124.5	0.172
	K14	0 - 5 cm	June	53,543,681	126.6	0.294
	K18	20 - 30 cm (E)	June	30,610,876	129.2	0.226
	K19	20 - 30 cm (M)	June	31,784,017	129.1	0.203
	K20	20 - 30 cm (W)	June	49,463,716	126.8	0.427
Urbana	K1	0 - 5 cm	November	21,681,291	124.5	0.101
	K3	20 - 30 cm	November	26,427,577	123.9	0.155
	K5	0 - 5 cm	September	28,018,675	123.7	0.188
	K7	20 - 30 cm	September	26,864,164	124.3	0.159
	K9	0 - 5 cm	April	32,535,582	124.6	0.127
	K11	20 - 30 cm	April	30,652,664	124.3	0.215
	K13	0 - 5 cm	June	34,023,870	124.4	0.162
	K15	20 - 30 cm (M)	June	29,187,308	127.0	0.237
	K16	20 - 30 cm (S)	June	72,914,672	126.9	0.492
	K17	20 - 30 cm (N)	June	32,057,255	126.0	0.466

Table C.4. Summary for co-assemblies from Havana and Urbana.

Samples	Depth	Million reads	IDBA co-assembly				Gene Prediction	
			Contigs	N50	Avg. length	Longest contig	Total bp	Genes
Havana top	0-5 cm	136,453,108	118,687	1,130	1,160.5	46,851	137,742,067	220,365
Havana deep	20-30 cm	179,133,698	419,023	1,779	1,568.9	388,680	657,447,015	938,759
Urbana top	0-5 cm	104,056,954	147,610	1,349	1,308.9	60,203	193,216,907	301,988
Urbana deep	20-30 cm	195,425,606	430,724	1,524	1,409.7	78,105	607,223,845	883,376

Table C.5. Summary of ROCker models used for detecting N genes in metagenomic soil samples.

Target Protein	ROCker build	
	Positive references	Negative references
AmoA bacteria	7	14
AmoA archaea	5	16
Hao	22	9
NarG/NxrA	311	0
NirK (Clade I and II)	140	0
NirK (Thaumarchaeota)	18	0
NirK (Nitrospira)	10	0
NirS	74	33
NorB	309	0
NosZ	166	0
RpoB	756	0

Table C.6. Summary for obtained bins from Havana and Urbana.

Co-assembly	Bin	Phyllum	Best classification level	Closest relative (AAI)	Novelty	Complet	Cont	Genome size	GC
Havana deep	HD001	*Thaumarchaeota*	**family** Nitrosopumilaceae (p-value: 0)	Candidatus Nitrosoarchaeum koreensis MY1 GCA 000220175 (79.74% AAI)	subspecies (p-value: 0)	96.2%	0.0%	1,618,161	0.36
	HD004	*Thaumarchaeota*	**family** of Candidatus Nitrosotenuis cloacae (p-value: 0.00082)	Candidatus Nitrosotenuis cloacae NZ CP011097 (63.26% AAI)	species (p-value: 0.0054)	92.3%	0%	1,433,284	0.36
	HD017	*Nitrospirae*	**order** *Nitrospirales* (p-value: 0.0099)	Nitrospira defluvii NC 014355 (56.81% AAI)	species (p-value: 0.0038)	80.2%	0.9%	3,655,691	0.56
	HD019	*Thaumarchaeota*	**class** Nitrososphaeria (p-value: 0)	Candidatus Nitrososphaera gargensis Ga9 2 CP002408 (56.03% AAI)	species (p-value: 0.0038)	92.3%	0%	1,872,034	0.36
	HD048	*Archaea*	**domain** *Archaea* (p-value: 0)	Aciduliprofundum sp MAR08 339 NC 019942 (39.0% AAI)	family (p-value: 0.0087)	46.2%	7.7%	1,174,600	0.65
	HD051	*Proteobacteria*	**phylum** *Proteobacteria* (p-value: 1.7e-05)	Geobacter bemidjiensis Bem NC 011146 (42.75% AAI)	genus (p-value: 0.0018)	67.6%	1.8%	5,714,526	0.57
	HD098	*Acidobacteria*	**phylum** *Acidobacteria* (p-value: 3.3e-05)	Terriglobus saanensis SP1PR4 NC 014963 (42.58% AAI)	genus (p-value: 0.0011)	52.3%	4.5%	3,731,758	0.57
	HD116	*Bacteroidetes*	**order** *Chitinophagales* (p-value: 0.0078)	Flavisolibacter sp LCS9 NZ CP011390 (57.51% AAI)	species (p-value: 0.0038)	80.2%	0.9%	3,591,707	0.39
	HD002	*Gemmatimonadetes*	**class** *Gemmatimonadetes* (p-value: 0.004)	Gemmatirosa kalamazoonesis CP007128 (46.18% AAI)	genus (p-value: 0.0087)	89.2%	0.9%	3,821,378	0.68
	HD025	*Proteobacteria*	**phylum** *Proteobacteria* (p-value: 0.046)	Pseudomonas aeruginosa CP014210 (39.76% AAI)	family (p-value: 0.0098)	71.2%	1.8%	3,331,498	0.67
	HD030	*Proteobacteria*	**phylum** *Proteobacteria* (p-value: 0.001)	Desulfuromonas soudanensis NZ CP010802 (41.74% AAI)	genus (p-value: 0)	70.3%	1.8%	6,048,977	0.71
	HD003	*Proteobacteria*	**phylum** *Proteobacteria* (p-value: 0.0092)	Pseudomonas aeruginosa CP014210 (40.72% AAI)	genus (p-value: 0)	78.4%	2.7%	3,648,130	0.69
	HD005	*Chloroflexi*	**phylum** *Chloroflexi* (p-value: 0.025)	Thermomicrobium roseum DSM 5159 NC 011959 (40.13% AAI)	genus (p-value: 0)	86.5%	0.9%	3,302,501	0.71
	HD009	*Proteobacteria*	**class** Alphaproteobacteria (p-value: 0)	Sphingomonas sp MM 1 NC 020561 (55.15% AAI	species (p-value: 0.0038)	84.7%	0.9%	2,124,243	0.66
	HD011	*Chloroflexi*	**phylum** *Chloroflexi* (p-value: 0.013)	Roseiflexus castenholzii DSM 13941 NC 009767 (40.57% AAI)	genus (p-value: 0)	57.7%	0.9%	1,736,863	0.70
	HD111	*Proteobacteria*	**phylum** *Proteobacteria* (p-value: 0.021)	Pseudomonas aeruginosa CP014210 (40.28% AAI)	genus (p-value: 0)	57.7%	0.9%	2,331,790	0.67
	HD113	*Acidobacteria*	**class** *Blastocatellia* (p-value: 0.0025)	Chloracidobacterium thermophilum B (46.58% AAI)	species (p-value: 0)	56.8%	4.5%	3,630,159	0.56
	HD006	*Proteobacteria*	**class** Alphaproteobacteria (p-value: 0)	Rhodospirillum centenum SW NC 011420 (50.14% AAI)	species (p-value: 0.00077)	74.8%	1.8%	4,798,044	0.64
	HD007	*Actinobacteria*	**phylum** *Actinobacteria* (p-value: 3.3e-05)	Thermomonospora curvata DSM 43183 NC 013510 (42.53% AAI)	genus (p-value: 0.0011)	79.3%	0.0%	2,271,926	0.69
	HD033	*Gemmatimonadetes*	**class** *Gemmatimonadetes* (p-value: 0.005)	Gemmatimonas aurantiaca T 27 (45.95% AAI)	genus (p-value: 0.0071)	83.8%	0.9%	4,059,537	0.67
	HD094	*Proteobacteria*	**phylum** *Proteobacteria* (p-value: 0.0015)	Geoalkalibacter subterraneus NZ CP010311 (41.65% AAI)	genus (p-value: 0)	52.3%	16.2%	2,233,560	0.65
	HD112	*Acidobacteria*	**class** Acidobacteriia (p-value: 0)	Terriglobus saanensis SP1PR4 NC 014963 (47.73% AAI)	species (p-value: 0.00077)	64.0%	2.7%	6,706,724	0.57
	HD008	*Chloroflexi*	**phylum** *Chloroflexi* (p-value: 0.018)	Roseiflexus castenholzii DSM 13941 NC 009767 (40.34% AAI)	genus (p-value: 0)	76.6%	0.9%	2,648,322	0.72
	HD010	*Actinobacteria*	**phylum** Actinobacteria (p-value: 0)	Conexibacter woesei DSM 14684 NC 013739 (43.01% AAI)	genus (p-value: 0.0021)	73.0%	0.9%	3,982,623	0.68
	HD054	*Actinobacteria*	**class** Actinobacteria (p-value: 0)	Ilumatobacter coccineus YM16 304 NC 020520 (43.93% AAI)	genus (p-value: 0.0026)	50.5%	12.6%	4,285,168	0.68
	HD012	*Proteobacteria*	**hylum** *Proteobacteria* (p-value: 0.025)	Desulfuromonas soudanensis NZ CP010802 (40.19% AAI)	genus (p-value: 0)	64.0%	1.8%	3,050,864	0.68
	HD013	*Proteobacteria*	**phylum** *Proteobacteria* (p-value: 0.0066)	Pseudomonas aeruginosa CP014210 (40.98% AAI)	genus (p-value: 0)	67.6%	3.6%	5,170,600	0.71
	HD022	*Actinobacteria*	**phylum** *Actinobacteria* (p-value: 0.00011)	Thermomonospora curvata DSM 43183 NC 013510 (42.35% AAI)	genus (p-value: 0.00053)	63.1%	3.6%	2,239,975	0.68
	HD067	*Firmicutes*	**domain** *Bacteria* (p-value: 0)	Pseudomonas aeruginosa CP014210 (39.08% AAI	family (p-value: 0.0087)	66.7%	0.9%	2,686,010	0.67
	HD103	*Acidobacteria*	**phylum** *Acidobacteria* (p-value: 3.3e-05)	Chloracidobacterium thermophilum B NC 016024 (42.54% AAI)	genus (p-value: 0.0011)	59.5%	2.7%	4,038,853	0.54
	HD020	*Proteobacteria*	**phylum** *Proteobacteria* (p-value: 0.0078)	Geoalkalibacter subterraneus NZ CP010311 (40.83% AAI)	genus (p-value: 0)	51.4%	0.9%	4,162,869	0.69
	HD073	*Acidobacteria*	**phylum** *Acidobacteria* (p-value: 0)	Terriglobus saanensis SP1PR4 NC 014963 (43.06% AAI)	genus (p-value: 0.0021)	63.1%	11.7%	3,113,876	0.62
	HD089	*Acidobacteria*	**phylum** *Acidobacteria* (p-value: 0.0015)	Candidatus Solibacter usitatus Ellin6076 NC 008536 (41.69% AAI)	genus (p-value: 0)	69.4%	3.6%	5,433,461	0.51
	HD021	*Nitrospirae*	**class** *Nitrospira* (p-value: 0)	Nitrospira defluvii NC 014355 (55.96% AAI)	species (p-value: 0.0038)	86.8%	1.8%	3,349,845	0.56
	HD028	*Acidobacteria*	**phylum** *Acidobacteria* (p-value: 0.0066)	Terriglobus saanensis SP1PR4 NC 014963 (40.98% AAI)	genus (p-value: 0)	80.2%	6.3%	6,230,186	0.63
	HD038	*Acidobacteria*	**class** Acidobacteria (p-value: 0)	Terriglobus saanensis SP1PR4 NC 014963 (48.18% AAI)	species (p-value: 0.00077)	90.1%	1.8%	6,335,714	0.58
	HD071	*Proteobacteria*	**phylum** *Proteobacteria* (p-value: 0.00071)	Geoalkalibacter subterraneus NZ CP010311 (41.84% AAI)	genus (p-value: 0)	76.6%	10.8%	6,698,088	0.69
	HD027	*Gemmatimonadetes*	**class** *Gemmatimonadetes* (p-value: 0.041)	Gemmatimonas aurantiaca T 27 (44.69% AAI)	genus (p-value: 0.0029)	51.4%	7.2%	4,054,031	0.70
	HD032	*Thaumarchaeota*	**class** Nitrososphaeria (p-value: 0)	Candidatus Nitrososphaera gargensis Ga9 2 CP002408 (54.48% AAI)	species (p-value: 0.0038)	88.5%	3.8%	2,630,146	0.35
	HD091	*Proteobacteria*	**class** Alphaproteobacteria (p-value: 0)	Blastochloris viridis NZ AP014854 (48.16% AAI)	species (p-value: 0.00077)	58.6%	0.9%	2,474,446	0.55
	HD102	*Firmicutes*	**phylum** *Firmicutes* (p-value: 0.025)	Alicyclobacillus acidocaldarius subsp acidocaldarius Tc 4 1 NC 017167 (40.15% AAI)	genus (p-value: 0)	57.7%	18.0%	1,655,952	0.53
	HD109	*Thaumarchaeota*	**family** of Candidatus Nitrosotenuis cloacae (p-value: 0)	Candidatus Nitrosotenuis cloacae NZ CP011097 (76.48% AAI)	subspecies (p-value: 0)	84.6%	0.0%	1,519,486	0.41
	HD034	*Proteobacteria*	**class** Betaproteobacteria (p-value: 0)	Nitrosospira multiformis ATCC 25196 NC 007614 (52.66% AAI)	species (p-value: 0.0015)	87.4%	9.0%	3,662,478	0.61
	HD082	*Acidobacteria*	**phylum** *Acidobacteria* (p-value: 0.00071)	Terriglobus saanensis SP1PR4 NC 014963 (41.84% AAI)	genus (p-value: 0)	53.2%	1.8%	3,785,730	0.60
Havana top	HT001	*Thaumarchaeota*	**class** Nitrososphaeria (p-value: 0)	Candidatus Nitrososphaera gargensis Ga9 2 CP002408 (56.13% AAI)	species (p-value: 0.0038	76.9%	19.2%	2,499,468	0.36
	HT003	*Proteobacteria*	**phylum** *Proteobacteria* (p-value: 0.034)	Pseudomonas aeruginosa CP014210 (39.98% AAI)	family (p-value: 0.0099)	76.6%	1.8%	2,908,538	0.71
	HT008	*Actinobacteria*	**phylum** Actinobacteria (p-value: 0)	Frankia inefficax NC 014666 (42.84% AAI)	genus (p-value: 0.0018)	57.7%	9.0%	1,960,640	0.67
Urbana deep	UD001	*Thaumarchaeota*	**class** Nitrososphaeria (p-value: 0)	Candidatus Nitrososphaera gargensis Ga9 2 CP002408 (56.12% AAI)	species (p-value: 0.0038)	92.3%	0.0%	1,941,192	0.36
	UD002	*Verrucomicrobia*	**class** Spartobacteria (p-value: 0.0027)	Candidatus Xiphinematobacter sp Idaho Grape NZ CP012665 (46.45% AAI)	species (p-value: 0)	66.7%	0.0%	2,597,474	0.55
	UD005	*Verrucomicrobia*	**phylum** Verrucomicrobia (p-value: 0)	Candidatus Xiphinematobacter sp Idaho Grape NZ CP012665 (45.1% AAI)	genus (p-value: 0.0034)	74.8%	15.3%	2,519,892	0.55
	UD010	*Actinobacteria*	**phylum** *Actinobacteria* (p-value: 8.1e-05)	Acidothermus cellulolyticus 11B NC 008578 (42.45% AAI)	genus (p-value: 0.00053)	54.1%	2.7%	2,443,042	0.68
	UD013	*Actinobacteria*	**phylum** *Actinobacteria* (p-value: 3.3e-05)	Acidothermus cellulolyticus 11B NC 008578 (42.56% AAI)	genus (p-value: 0.0011)	36.9%	5.4%	3,751,106	0.70
	UD015	*Proteobacteria*	**class** Deltaproteobacteria (p-value: 0)	Stigmatella aurantiaca DW4 3 1 NC 014623 (55.19% AAI)	species (p-value: 0.0038)	82.0%	3.6%	4,098,408	0.68
	UD003	*Actinobacteria*	**phylum** *Actinobacteria* (p-value: 0.00048)	Blastococcus saxobsidens DD2 NC 016943 (41.94% AAI)	genus (p-value: 0)	59.5%	4.5%	2,086,766	0.70
	UD007	*Verrucomicrobia*	**phylum** Verrucomicrobia (p-value: 0)	Candidatus Xiphinematobacter sp Idaho Grape NZ CP012665 (45.45% AAI)	genus (p-value: 0.0045)	67.6%	18.0%	3,066,617	0.55
	UD022	*Acidobacteria*	**class** Acidobacteriia (p-value: 0)	Terriglobus saanensis SP1PR4 NC 014963 (49.17% AAI)	species (p-value: 0.00077)	55.0%	2.7%	3,554,519	0.58
	UD035	*Actinobacteria*	**phylum** *Acidobacteria* (p-value: 0.00071)	Chloracidobacterium thermophilum B NC 016024 (41.85% AAI)	genus (p-value: 0)	56.8%	19.8%	5,510,755	0.66
	UD029	*Firmicutes*	**phylum** *Proteobacteria* (p-value: 0.046)	Pseudomonas aeruginosa CP014210 (39.8% AAI)	family (p-value: 0.0098)	60.4%	2.7%	3,408,806	0.66
	UD045	*Acidobacteria*	**phylum** *Acidobacteria* (p-value: 1.7e-05)	Chloracidobacterium thermophilum B NC 016024 (42.78% AAI)	genus (p-value: 0.0018)	82.0%	6.3%	4,569,695	0.56
	UD050	*Chloroflexi*	**phylum** *Chloroflexi* (p-value: 0.046)	Sphaerobacter thermophilus DSM 20745 NC 013523 (39.71% AAI)	family (p-value: 0.0098)	88.3%	11.7%	5,687,437	0.52
	UD053	*Acidobacteria*	**phylum** Acidobacteria (p-value: 0)	Chloracidobacterium thermophilum B NC 016024 (42.98% AAI)	genus (p-value: 0.0021)	74.8%	4.5%	5,401,512	0.54
	UD057	*Proteobacteria*	**phylum** *Proteobacteria* (p-value: 0.001)	Pseudomonas aeruginosa CP014210 (41.75% AAI)	genus (p-value: 0)	54.1%	2.7%	4,060,224	0.71
	UD059	*Deinococcus-Thermus**	**phylum** Deinococcus-Thermus (p-value: 0.0056)	Thermus brockianus CP016312 (41.09% AAI)	genus (p-value: 0)	48.6%	15.3%	5,643,751	0.71
	UD061	*Proteobacteria*	**class** Alphaproteobacteria (p-value: 0)	Rhodospirillum centenum SW NC 011420 (48.32% AAI)	species (p-value: 0.00077)	67.6%	4.5%	4,656,205	0.67
	UD063	*Nitrospirae*	**class** *Nitrospira* (p-value: 0)	Nitrospira defluvii NC 014355 (55.9% AAI)	species (p-value: 0.0038)	82.0%	5.4%	3,764,772	0.57
Urbana Top	UT003	*Bacteria*	**domain** *Bacteria* (p-value: 0)	Pseudomonas aeruginosa CP014210 (39.59% AAI)	family (p-value: 0.0094)	55.0%	9.9%	3,304,733	0.71
	UT008	*Proteobacteria*	**class** Alphaproteobacteria (p-value: 0)	Azospirillum humicireducens NZ CP015285 (50.71% AAI)	species (p-value: 0.0015)	62.2%	1.8%	3,440,494	0.65
	UT009	*Acidobacteria*	**class** Acidobacteriia (p-value: 0)	Terriglobus saanensis SP1PR4 NC 014963 (48.56% AAI)	species (p-value: 0.00077)	63.1%	4.5%	3,728,385	0.57
	UT011	*Verrucomicrobia*	**class** *Spartobacteria* (p-value: 0.005)	Candidatus Xiphinematobacter sp Idaho Grape NZ CP012665 (45.95% AAI)	genus (p-value: 0.0071)	65.8%	0.0%	2,550,761	0.54

Table C.7. Summary of Glycoside hydrolase enzymes found in metagenomic bins

	Bin	Cellulase						Endohemicellulases				Cell wall elongation		Debranching enzymes		Oligosaccharide-degrading enzymes										Not classified	Amylolytic enzymes					Other degradative enzymes					
		GH5	GH6	GH9	GH44	GH1	GH5	GH94	GH10	GH26	GH28	GH63	GH78	GH16	GH74	GH51	GH67	GH2	GH3	GH29	GH35	GH38	GH39	GH42	GH43	GH65	GH31	GH4	GH27	GH0	GH13	GH15	GH57	GH13	GH77	GH23	GH11	GH20	GH18	GH19	GH13
Urbana deep	UD001	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	UD002	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	4	0	0	0	1	0	0	0	0	0	0
	UD003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	0	0	0	0	0	0	0	0	0
	UD005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4	0	0	0	0	0	0	0	0	0	0
	UD007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3	5	2	0	0	0	0	0	0	0	0	0
	UD010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	0	0	0	3	0	0	0	1	0	0	0	0	0
	UD013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	6	1	0	0	0	0	0	0	1	0	0
	UD015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3	5	0	0	0	1	0	0	0	0	0	0
	UD022	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	7	1	1	0	1	0	0	0	0	0	0
	UD029	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	8	0	0	0	0	0	2	0	0	0	0	4	7	0	0	0	0	0	0	0	0	0	3
	UD035	1	0	0	0	0	0	0	0	0	0	0	0	11	0	3	0	0	3	0	0	0	0	0	3	0	0	1	0	15	14	0	0	0	0	1	0	0	0	1	2
	UD045	1	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	6	8	0	0	0	0	0	0	0	0	1	0	6	1	0	0	0	0	1	0	0	0	0	0
	UD050	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	0	0	0	0	0	3	1	0	3	0	1	9	0	0	0	1	0	0	0	0	0	1
	UD053	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	5	0	0	0	0	0	2	1	0	3	0	4	2	0	0	0	1	0	0	0	1	0	0
	UD057	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	1	0	0	0	0	4	2	0	0	0	1	0	0	0	0	0	0
	UD059	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	9	2	0	0	0	0	0	0	0	0	0
	UD061	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4	3	0	0	0	0	0	0	0	0	0
	UD063	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	8	3	2	1	0	1	1	0	0	0	1	0
Urbana top	UT003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	1	0	0	1	0	0	0	1	0	0
	UT008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	7	3	0	1	0	1	0	0	1	0	0
	UT009	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	7	1	0	0	0	0	0	0	1	0	2
	UT011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	6	1	0	0	0	0	0	0	1	0	0
Havana deep	HD001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
	HD002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
	HD003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	HD004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	HD005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
	HD006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	5	1	0	0	0	0	0	0	0	0	0
	HD007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0
	HD008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	0	0	0	0	0	0	0	0	0
	HD009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
	HD010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	3	0	0	0	0	0	0	0	0	0
	HD011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	HD012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1	0	0	0	3	1	0
	HD013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	8	1	0	0	0	0	0	0	0	0	0
	HD017	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	4	9	0	1	1	1	6	0	0	0	0	0
	HD019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	HD020	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	4	0	1	2	0	0	0	0	0	0	0
	HD021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	11	2	1	0	1	0	6	0	0	0	0	2
	HD022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
	HD025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	HD027	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	0	0	0	1	0	0	0	0	0	1
	HD028	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	4	5	2	0	0	0	0	0	0	1	1	1
	HD030	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	0	0	0	0	0	0	0	0	0	1
	HD032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	HD033	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	4	7	1	0	1	1	0	0	0	0	0	0
	HD034	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
	HD038	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	5	3	2	1	0	0	0	0	0	0	0	1
	HD048	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	HD051	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	1
	HD054	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	4	2	0	0	4	0	3	0	0	0	0	0	0	0
	HD067	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3	6	2	0	0	0	0	0	0	0	0	0
	HD071	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	9	2	0	0	0	0	0	0	0	0	0
	HD073	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	2	0	1
	HD082	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	1	0	0	0	0	0	0
	HD089	2	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	1	0	0	0	0	1
	HD091	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	HD094	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	HD098	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	1	1	1	1	0	0	0	2	1	0
	HD102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
	HD103	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	1	0	0	0	0	2	0
	HD109	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	HD111	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	1
	HD112	2	0	2	0	0	0	0	0	0	0	1	0	0	0	4	1	4	3	3	1	0	0	2	0	0	3	0	2	6	2	5	1	0	2	1	2	0	0	1	1
	HD113	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	2	2	0	1	0	0	1	0	0	0	1	1
Havana top	HT001	0	0	1	0	0	0	0	1	0	0	2	0	0	0	3	1	0	3	2	1	0	0	4	0	0	0	0	0	5	4	0	0	0	2	0	0	0	0	0	1
	HT003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	5	1	0	0	0	0	0	0	0	0	0
	HT008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	2	5	2	0	0	0	0	0	0	0	0	0

APPENDIX D: SUPPLEMENTARY MATERIAL FOR CHAPTER 5

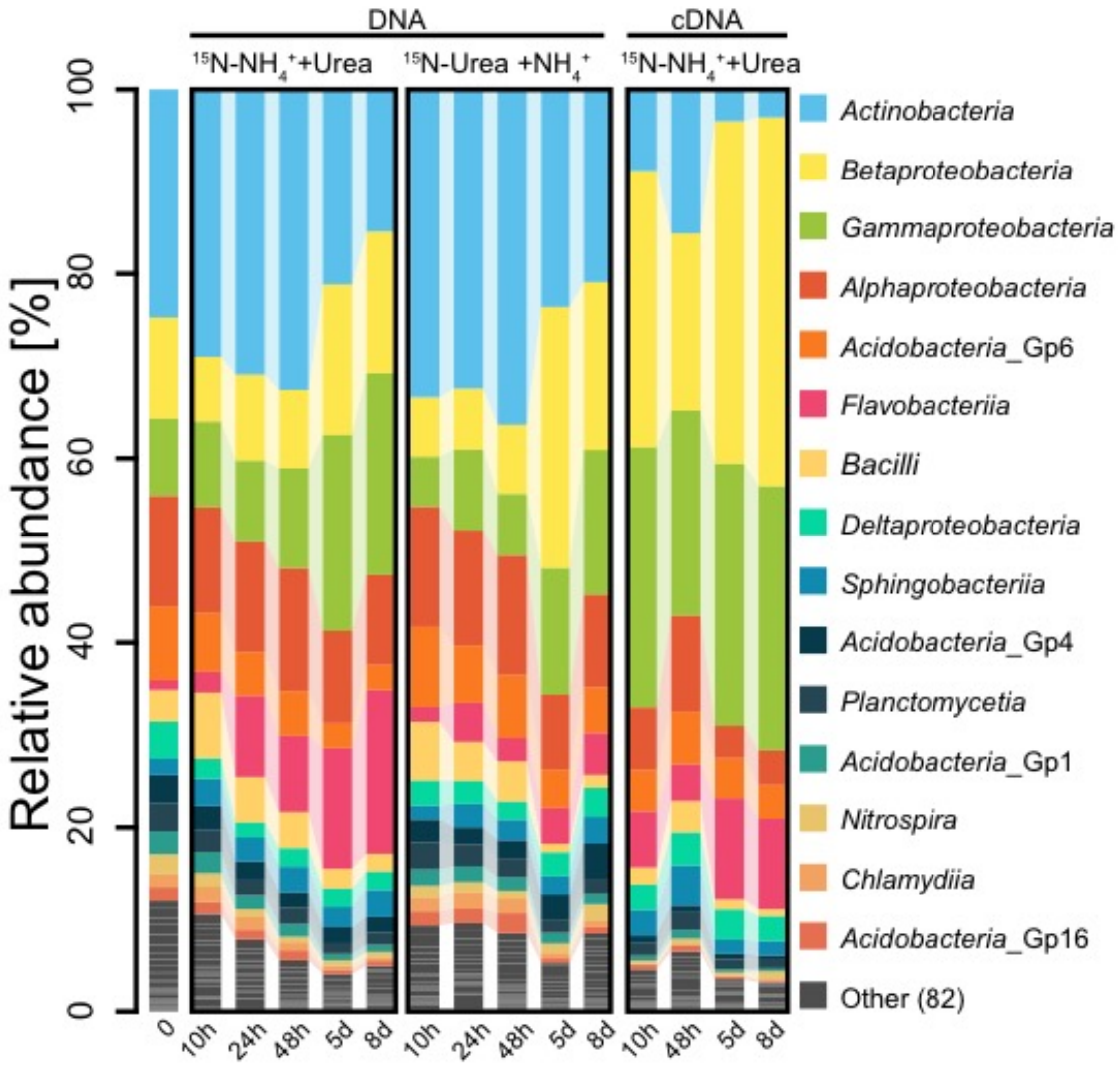


Figure D.1. Taxonomic composition of soil incubations. Summary of the taxonomic affiliation of recovered from 16S rRNA gene and transcript fragments from metagenomes and metatranscriptomes.

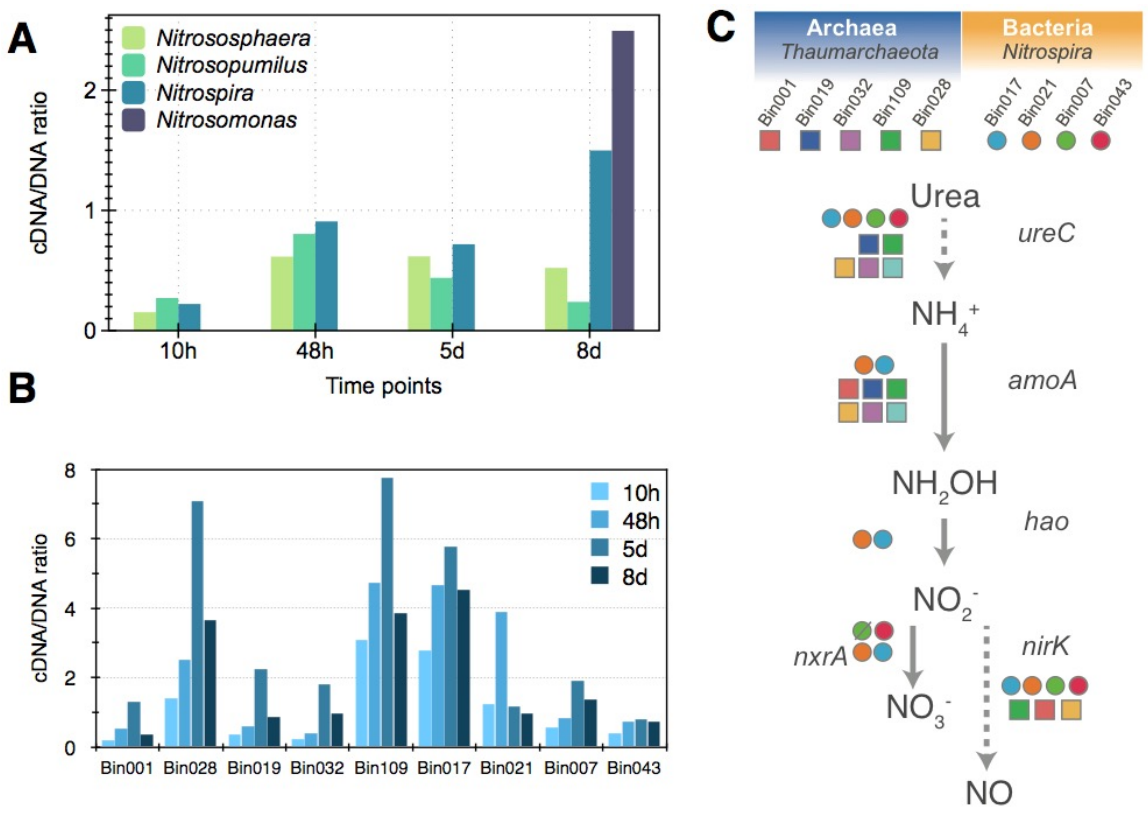


Figure D.2. Abundance of indigenous ammonia-oxidizing archaea, ammonia-oxidizing bacteria and nitrite-oxidizing bacteria population genomes in metagenomes and metatranscriptomes. A. Relative expression ratio (cDNA/DNA) for 16S rRNA gene fragments related to nitrifier communities (genus level). B. Relative expression values for nitrifier bins obtained from incubation metagenomes and previous field metagenomes are displayed from each incubation point. C. Metagenomic bins obtained from soil incubations were queried for the presence of hallmark nitrification gene markers using HMM models. Each shape represents an individual metagenomic bin. Arrows show predicted nitrogen cycle pathways.

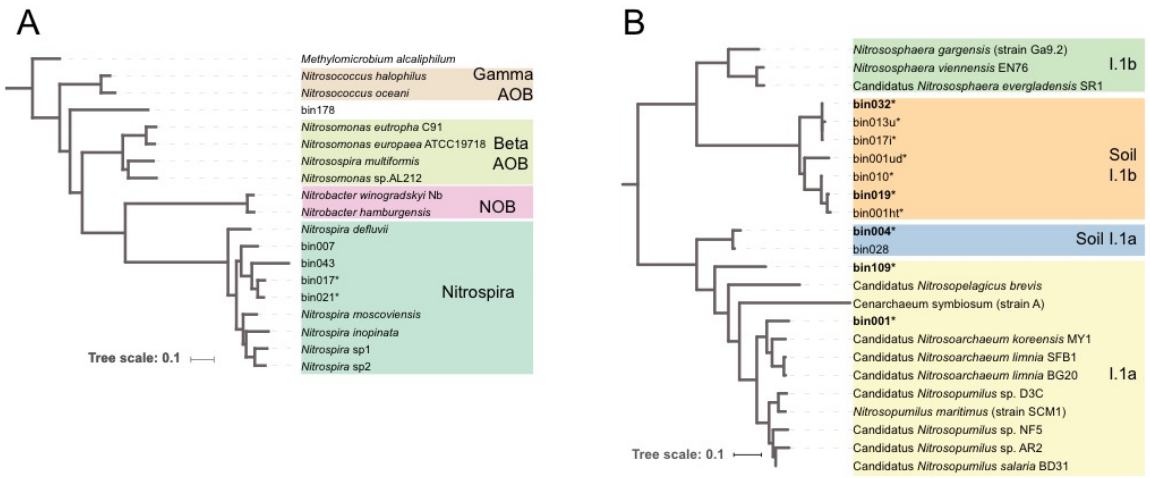


Figure D.3. Recovery of indigenous archaeal and bacterial ammonia-oxidizing populations in incubated soils. Phylogenetic reconstruction of bacterial (A) and archaeal (B) genomic bins potentially participating in nitrification. Concatenated alignments of conserved genes for bacterial or archaeal genomes were used to build maximum likelihood trees in RAxML. Bins reported previously are marked with an asterisk symbol.

Table D.1 Moisture and pH for incubated soils.

Condition	Time point	pH			Moisture		
Control	0h	7.42	±	0.09	7.88	±	0.49
	10h	7.51	±	0.06	8.01	±	0.11
	24h	7.49	±	0.06	8.43	±	0.07
	48h	7.38	±	0.23	8.49	±	0.25
	5d	7.48	±	0.05	7.82	±	0.40
	8d	7.54	±	0.03	8.10	±	0.12
N-ammended	0h	7.53	±	0.02	7.87	±	0.39
	10h	7.61	±	0.03	7.86	±	0.24
	24h	7.64	±	0.05	7.94	±	0.42
	48h	7.59	±	0.02	7.77	±	0.24
	5d	7.50	±	0.04	7.71	±	0.21
	8d	6.60	±	0.13	7.92	±	0.15

Table D.2. Soil incubation metagenome statistics

ID	Test	Time point	Triming		Nonpareil
			Coupled reads	Average read length	Coverage
Cont_0	Control	0*	53,481,680	140.9	0.424
Cont_5d		5d*	23,780,198	140.1	0.308
Cont_8d		8d*	32,122,126	141.8	0.369
Amo_10h	^{15}N NH_4^+ + Urea	10h	12,897,828	140.6	0.356
Amo_24h		24h	8,221,766	136.6	0.497
Amo_48h		48h	25,938,078	136.5	0.425
Amo_5d		5d*	24,204,426	141.5	0.294
Amo_8d		8d*	36,660,412	139	0.347
Urea_10h	NH_4^+ + ^{15}N Urea	10h	18,721,684	140.5	0.311
Urea_24h		24h	15,367,868	134.7	0.431
Urea_48h		48h	13,840,496	140.5	0.283
Urea_5d		5d	28,976,856	141.6	0.356
Urea_8d		8d	32,139,060	142.3	0.274

Table D.3. Soil incubation metatranscriptome statistics

ID	Test	Time point	Merging	Trimming			
			Percentage merged	Merged reads	Average read length	Not merged reads	Average read length
Cont_24h	Control	1d	85.5	21,167,418	138.6	4,192,488	140.7
Cont_5d		5d	82.5	16,126,520	154.4	4,431,284	141
Cont_8d		8d	74.5	14,592,479	151.5	6,534,598	141.1
Amo_10h	^{15}N NH_4^+ + Urea	10h	79.0	23,252,114	154.3	8,101,024	140.7
Amo_24h		24h	86.9	10,695,118	141	2,149,852	140.5
Amo_48h		48h	80.3	23,628,975	150.9	7,233,008	140.9
Amo_5d		5d	85.0	8,305,590	143.5	1,829,636	140.1
Amo_8d		8d	76.3	13,392,531	152.2	5,789,762	139.7

Table D.4. Ribosomal gene fragments recovered from metatranscriptomes

ID	rRNA											Non rRNA	
	Bacterial 16S	Archaea 16S	5.8S	5S	Achaea 23S	Bacteria 23S	Eukarya 18S	Eukarya 28S	Total [%]	16S/23S ratio	Sequences	Non RNA [%]	Sequences
Cont_24h	31.71	0.09	0.05	0.06	0.20	61.87	0.53	0.86	95.37	1.95	25,359,906	4.6	1,174,242
Cont_5d	35.10	0.09	0.08	0.00	0.23	58.21	0.73	1.07	95.51	1.66	20,557,804	4.5	922,754
Cont_8d	30.18	0.09	0.08	0.00	0.23	63.05	0.56	0.90	95.09	2.09	21,127,077	4.9	1,040,092
Amo_10h	30.76	0.10	0.10	0.04	0.22	57.89	2.06	3.96	95.13	1.88	31,353,138	4.9	1,529,334
Amo_48h	30.59	0.15	0.14	0.05	0.35	55.63	2.90	4.81	94.62	1.82	30,861,983	5.4	1,656,713
Amo_5d	35.91	0.09	0.07	0	0.22	60.37	0.55	0.72	97.93	1.68	10,135,226	2.1	1,719,081
Amo_8d	33.21	0.09	0.07	0	0.24	62.85	0.71	0.95	98.12	1.89	19,182,293	1.9	887,882

Table D.5. Referential soil microorganisms used as part of the database for spectra annotation.

Reference proteome	
Anaeromyxobacter dehalogenans 2CP-1	*Nocardia farcinica* (strain IFM 10152)
Anaeromyxobacter sp. Fw109-5	*Mycobacterium* sp. (strain KMS)
Gemmatimonas aurantiaca	*Sorangium cellulosum* (strain So ce56)
Gemmatirosa kalamazoonensis	*Paracoccus denitrificans* (strain Pd 1222)
Bradyrhizobium japonicum USDA	Thaumarchaeota archaeon N4
Dyadobacter fermentans	*Nitrosospira multiformis* (strain ATCC 25196)
Nitrososphaera gargensis (strain Ga9.2)	*Gaiella* sp. SCGC AG-212-M14
Candidatus *Nitrosopumilus* sp. AR2	*Flavobacterium johnsoniae*
Candidatus *Nitrosoarchaeum koreensis* MY1	Candidatus *Saccharibacteria* bacterium GW2011_GWC2_48_9
Candidatus *Nitrospira inopinata*	*Aeromicrobium* sp. Root344
Nitrosomonas europaea	*Pseudomonas fluorescens* (strain Pf0-1)
Nitrospira moscoviensis	*Koribacter versatilis* (strain Ellin345)
Candidatus *Nitrospira nitrosa*	*Acidobacterium capsulatum* (strain ATCC 51196)
Conexibacter woesei (strain DSM 14684)	*Marmoricola* sp. Leaf446
Nocardioides sp. (strain BAA-499 / JS614)	*Nocardioides* sp. Soil777
Pseudogulbenkiania sp. (strain NH8B)	*Sphingomonas sanxanigenens* DSM 19645
Rhodoferax ferrireducens (strain ATCC BAA-621)	*Solirubrobacter soli*
Acidovorax avenae (strain ATCC 19860)	*Rhodopseudomonas palustris* (strain ATCC BAA-98)
Phenylobacterium zucineum (strain HLK1)	
Oligotropha carboxidovorans (strain ATCC 49405)	
Hyphomicrobium denitrificans (strain ATCC 51888)	
Ralstonia pickettii (strain 12D)	
Leptothrix cholodnii (strain ATCC 51168)	
Actinosynnema mirum (strain ATCC 29888)	
Geobacillus kaustophilus (strain HTA426)	
Dechloromonas aromatica (strain RCB)	
Accumulibacter phosphatis (strain UW-1)	
Thioalkalivibrio nitratireducens (strain DSM 14787	
Frankia alni (strain ACN14a)	

Table D.6. Taxonomy and statistics for recovered bins

Bin	Complet.	Cont.	Best classification level	Novelty	Closest relative (AAI)
Bin001	76.6%	2.7%	**domain** Bacteria 0, **phylum** Actinobacteria 0.0459,	**family**(p value: 0.0098)	Geodermatophilus obscurus DSM 43160 NC 013757 (39.8% AAI)
Bin002	84.7 %.	3.6 %.	**domain** Bacteria 0,	**family** (p value: 0.0094)	Mycobacterium tuberculosis CP008969 (39.53% AAI)
Bin003	72.1 %.	0.9 %.	Bacteria 0, **phylum** Proteobacteria 0.0151	**genus** (p value: 0)	Pseudomonas aeruginosa CP014210 (40.4% AAI)
Bin004	91.9 %.	2.7%	Bacteria 0, **phylum** Acidobacteria 0.00196	**genus** (p value: 0)	Candidatus Solibacter usitatus Ellin6076 NC 008536 (41.55% AAI)
Bin005	66.7%	5.4%	Bacteria 0, **phylum** Proteobacteria 0.0248	**genus** (p value: 0)	Pseudomonas aeruginosa CP014210 (40.14% AAI)
Bin006	91.9%	3.6%	Bacteria 0, **phylum** Acidobacteria 0, **class** Acidobacteriia 0	**species** (p value: 0.0031)	Candidatus Koribacter versatilis Ellin345 NC 008009 (53.82% AAI)
Bin007	86.5%	0.9%	**domain** Bacteria 0, **phylum** Nitrospirae 0, **class** Nitrospira 0, **order** Nitrospirales 0.0111, **family** Nitrospiraceae 0.0431	**species** (p value: 0.0038)	Nitrospira defluvii NC 014355 (56.66% AAI)
Bin008	68.5%	9.9%	**domain** Bacteria 0, **phylum** Proteobacteria 0.0211	**genus** (p value: 0)	Pseudomonas aeruginosa CP014210 (40.26% AAI)
Bin011	72.1%	23.4%	Bacteria 0, **phylum** Proteobacteria 0.0179	**genus** (p value: 0)	Pseudomonas aeruginosa CP014210 (40.34% AAI)
Bin012	80.2%	0.9%	**domain** Bacteria 0, **phylum** Acidobacteria 0	**genus** (p value: 0.0021)	Chloracidobacterium thermophilum B NC 016024 (43.22% AAI)
Bin013	80.2%	21.6%	**domain** Bacteria 0, **phylum** Firmicutes 0.00297,	**genus** (p value: 0)	Desulfotomaculum kuznetsovii DSM 6115 CP002770 (41.32% AAI)
Bin015	76.6%	7.2%	**domain** Bacteria 0, **phylum** Acidobacteria 0.00235	**genus** (p value: 0)	Terriglobus saanensis SP1PR4 NC 014963 (41.43% AAI)
Bin015	76.6%	7.2%	**domain** Bacteria 0, **phylum** Acidobacteria 0.00235	**genus** (p value: 0)	Terriglobus saanensis SP1PR4 NC 014963 (41.43% AAI)
Bin018	84.7%	33.3%	**domain** Bacteria 0, **phylum** Proteobacteria 0.0211	**genus** (p value: 0)	Pseudomonas aeruginosa CP014210 (40.26% AAI)
Bin020	80.2%	19.8%	**domain** Bacteria 0, **phylum** Actinobacteria 3.34e 05	**genus** (p value: 0.0013)	Conexibacter woesei DSM 14684 NC 013739 (42.61% AAI)
Bin021	75.7%	7.2%	**domain** Bacteria 0, **phylum** Proteobacteria 0.0211,	**genus** (p value: 0)	Pseudomonas aeruginosa CP014210 (40.24% AAI)
Bin025	56.8%	35.1%	**domain** Bacteria 0, **phylum** Proteobacteria 0.00664	**genus** (p value: 0)	Pseudomonas aeruginosa CP014210 (40.93% AAI)
Bin026	74.8%	24.3%	**domain** Bacteria 0, **phylum** Proteobacteria 0, **class** Betaproteobacteria 0	**species** (p value: 0.00077)	Azoarcus sp BH72 NC 008702 (48.64% AAI)
Bin028	27.9%*	14.4%	**domain** Archaea 0, **phylum** Thaumarchaeota 0, **class** 0, **order** Nitrosopumilales 0, **family** Nitrosopumilaceae 0	**subspecies** (p value: 0)	Candidatus Nitrosoarchaeum limnia SFB1 CM001158 (70.7% AAI)
Bin031	87.4%	18%	**domain** Bacteria 0, **phylum** Acidobacteria 0	**genus** (p value: 0.0021)	Terriglobus saanensis SP1PR4 NC 014963 (42.93% AAI)
Bin033	61.3%	23.4%	**domain** Bacteria 0, **phylum** Proteobacteria 0.0211	**genus** (p value: 0)	Pseudomonas aeruginosa CP014210 (40.27% AAI)
Bin034	69.4%	13.5%	**domain** Bacteria 0, **phylum** Chloroflexi 0.0392	**family** (p value: 0.0099)	Roseiflexus castenholzii DSM 13941 NC 009767 (39.83% AAI)
Bin035	46.8%	35.1%	**domain** Archaea 0, **phylum** Euryarchaeota 0.0131	**genus** (p value: 0)	Aciduliprofundum sp MAR08 339 NC 019942 (40.58% AAI)
Bin043	58.6%	34.2%	**domain** Bacteria 0, **phylum** Nitrospirae 0, **class** Nitrospira 0,	**species** (p value: 0.0015)	Nitrospira defluvii NC 014355 (51.53% AAI)
Bin049	74.8%	13.5%	**domain** Bacteria 0, **phylum** Acidobacteria 0.000345	**genus** (p value: 0.00053)	Chloracidobacterium thermophilum B NC 016024 (42.08% AAI)
Bin072	64.9%	55.9%	**domain** Bacteria 0, **phylum** Acidobacteria 0.00555	**genus** (p value: 0)	Chloracidobacterium thermophilum B NC 016024 (41.1% AAI)
Bin081	66.7%	35.1%	**domain** Bacteria 0, **phylum** Proteobacteria 0.0337	**family** (p value: 0.0099)	Pseudomonas aeruginosa CP014210 (39.95% AAI)
Bin084	56.8%	61.3%	**domain** Bacteria 0, **phylum** Proteobacteria 0.0289	**genus** (p value: 0)	Pseudomonas aeruginosa CP014210 (40.02% AAI)
Bin095	80.2%	20.7%	**domain** Bacteria 0, **phylum** Acidobacteria 0	**genus** (p value: 0.0026)	Chloracidobacterium thermophilum B NC 016024 (43.92% AAI)
Bin097	82.9%	2.7%	**domain** Bacteria 0, **phylum** Proteobacteria 0, **class** Gammaproteobacteria 0	**species** (p value: 0.0031)	Acinetobacter lwoffii WJ10621 NZ CM001194 (53.74% AAI)
Bin102	77.5%	34.2%	**domain** Bacteria 0, **phylum** Proteobacteria 0, **class** Alphaproteobacteria 0	**species** (p value: 0.00077)	Starkeya novella DSM 506 NC 014217 (47.93% AAI)
Bin105	73%	30.6%	**domain** Bacteria 0, **phylum** Bacteroidetes 0, **class** Chitinophagia 0,	**species** (p value: 0.0038)	Niabella ginsenosidivorans NZ CP015772 (55.61% AAI)
Bin107	82%	9 %.	**domain** Bacteria 0, **phylum** Bacteroidetes 0, **class** Flavobacteriia 0, **order** Flavobacteriales 0, **family** Crocinitomicaceae 0.00059	**species** (p value: 0.0054)	Fluviicola taffensis DSM 16823 NC 015321 (64.27% AAI)
Bin111	55.9%	2.7%	**domain** Bacteria 0, **phylum** Bacteroidetes 0, **class** Flavobacteriia 0, **order** Flavobacteriales 0, **family** Flavobacteriaceae 0.000609	**species** (p value: 0.0054)	Flavobacterium columnare NZ CP015107 (64.51% AAI)
Bin121	66.7%	49.5 %.	**domain** Bacteria 0, **phylum** Bacteroidetes 0, **class** Flavobacteriia 0.00396	**genus** (p value: 0.0087)	Candidatus Sulcia muelleri CP016223 (46.1% AAI)
Bin134	81.1%	31.5%	**domain** Bacteria 0, **phylum** Candidatus Saccharibacteria 0, **class** 0, **order** 0.313, **family** 0.446	**species** (p value: 0.0015)	Candidatus Saccharibacteria bacterium GW2011 GWC2 44 17 CP011211 (50.97% AAI)
Bin135	73%	59.5%	**domain** Bacteria 0, **phylum** Actinobacteria 0, **class** Actinobacteria 0, **order** Propionibacteriales 0.00517, **family** Nocardioidaceae 0.0192	**species** (p value: 0.0038)	Nocardioides sp JS614 NC 008699 (57.76% AAI)
Bin142	49.5%	5.4%	**domain** Bacteria 0, **phylum** Proteobacteria 0, **class** Gammaproteobacteria 0, **order** Xanthomonadales 0, **family** Rhodanobacteraceae 0, **genus** Rhodanobacter 0.0107	**subspecies** (p value: 0)	Rhodanobacter denitrificans NC 020541 (87.39% AAI)
Bin144	73%	10.8%	**domain** Bacteria 0, **phylum** Actinobacteria 0, **class** Actinobacteria 0, **order** Propionibacteriales 0.000223, **family** Nocardioidaceae 0.0105	**species** (p value: 0.0038)	Nocardioides sp JS614 NC 008699 (58.18% AAI)
Bin146	72.1%	7.2%	**domain** Bacteria 0, **phylum** Proteobacteria 0, **class** Alphaproteobacteria 0, **order** Sphingomonadales 0, **family** Sphingomonadaceae 0, **genus** Sphingopyxis 0.0153	**subspecies** (p value: 0)	Sphingopyxis fribergensis NZ CP009122 (85.97% AAI)
Bin151	80.2%	32.4 %%	**domain** Bacteria 0, **phylum** Bacteroidetes 0, **class** Flavobacteriia 0, **order** Flavobacteriales 0, **family** Flavobacteriaceae 0	**species** (p value: 0.0085)	Flavobacterium columnare NZ CP015107 (67.17% AAI)
Bin152	47.7%	40.5%	**domain** Bacteria 0, **phylum** Actinobacteria 0, **class** Actinobacteria 0,	**species** (p value: 0.00077)	Mycobacterium fortuitum subsp fortuitum DSM 46621 ATCC 6841 CP014258 (48.45% AAI)
Bin159	77.5%	23.4%	**domain** Bacteria 0, **phylum** Actinobacteria 0, **class** Actinobacteria 0, **order** Propionibacteriales 0.0102, **family** Nocardioidaceae 0.0387	**species** (p value: 0.0038)	Aeromicrobium erythreum NZ CP011502 (56.75% AAI)
Bin164	87.4%	67.6%	**domain** Bacteria 0, **phylum** Proteobacteria 0, **class** Deltaproteobacteria 0	**species** (p value: 0.0015)	Bdellovibrio bacteriovorus W NZ CP002190 (52.99% AAI)
Bin167	82.9%	4.5%	**domain** Bacteria 0, **phylum** Proteobacteria 0, **class** Gammaproteobacteria 0	**species** (p value: 0.0038)	Lysobacter capsici NZ CP011130 (55.14% AAI)
Bin176	92.8%	1.8%	**domain** Bacteria 0, **phylum** Proteobacteria 0, **class** Betaproteobacteria 0, **order** Burkholderiales 0, **family** Oxalobacteraceae 0	**species** (p value: 0.0062)	Janthinobacterium sp Marseille NC 009659 (66.24% AAI)
Bin175	92.8%	27%	**domain** Bacteria 0, **phylum** Proteobacteria 0, **class** Deltaproteobacteria 0	**species** (p value: 0.0031),	Bdellovibrio exovorus JSS NC 020813 (53.9% AAI)
Bin178	95.5%	0.9%	**domain** Bacteria 0, **phylum** Proteobacteria 0, **class** Gammaproteobacteria 0.00568	**genus** (p value: 0.0063)	Thioalkalivibrio sulfidiphilus HL EbGr7 NC 011901 (45.83% AAI)

Table D.7. Annotations for most abundant detected peptides in control and N-amended incubations.

Incubation	Abundance Ranking	Average SpC	Predicted function	Best match organisms
Control	1	724.6	Chaperone protein DnaK	Methylococcus capsulatus (strain ATCC 33009 / NCIMB 11132 / Bath)
	2	403.63	Elongation factor Tu	Rhodopseudomonas palustris (strain BisB18)
	3	361.75	Flagellin	Leptothrix cholodnii (strain ATCC 51168 / LMG 8142 / SP 6)
	4	329.57	Methanol dehydrogenase large subunit like protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	5	313.55	ATP dependent Clp protease ATP binding subunit ClpX	Gaiella sp. SCGC AG 212 M14
	6	299.14	Bll6026 protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	7	294.02	Bll6025 protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	8	285.42	ATP dependent Clp protease ATP binding subunit ClpX	Caldanaerobacter subterraneus subsp. tengcongensis (strain DSM 15242 / JCM 11007 / NBRC 100824 / MB4)
	9	283.48	Bll6026 protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	10	283.48	Putative outer membrane protein B, OmpB	Koribacter versatilis (strain Ellin345)
	11	270.87	ATP dependent Clp protease ATP binding subunit ClpX	Koribacter versatilis (strain Ellin345)
	12	263.17	Elongation factor Tu	Burkholderia lata (strain ATCC 17760 / LMG 22485 / NCIMB 9086 / R18194 / 383)
	13	239.04	Formaldehyde activating enzyme	Hyphomicrobium denitrificans (strain ATCC 51888 / DSM 1869 / NCIB 11706 / TK 0415)
	14	227.23	Bll6025 protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	15	224.86	Chaperone protein DnaK	Azorhizobium caulinodans (strain ATCC 43989 / DSM 5975 / JCM 20966 / NBRC 14845 / NCIMB 13405 / ORS
	16	219.46	ABC transporter ATP binding protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	17	216.17	Bll6026 protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	18	213.1	Formaldehyde activating enzyme	Hyphomicrobium denitrificans (strain ATCC 51888 / DSM 1869 / NCIB 11706 / TK 0415)
	19	212.75	Phasin	Dechloromonas aromatica (strain RCB)
	20	212.45	Elongation factor Tu	Dechloromonas aromatica (strain RCB)
NH_4^+ + Urea	1	1132.34	Formaldehyde activating enzyme	Hyphomicrobium denitrificans (strain ATCC 51888 / DSM 1869 / NCIB 11706 / TK 0415)
	2	1132.34	Formaldehyde activating enzyme	Hyphomicrobium denitrificans (strain ATCC 51888 / DSM 1869 / NCIB 11706 / TK 0415)
	3	983.89	Formaldehyde activating enzyme	Hyphomicrobium denitrificans (strain ATCC 51888 / DSM 1869 / NCIB 11706 / TK 0415)
	4	422.18	PQQ dependent dehydrogenase, methanol/ethanol family	Hyphomicrobium denitrificans (strain ATCC 51888 / DSM 1869 / NCIB 11706 / TK 0415)
	5	404.26	Methanol dehydrogenase large subunit like protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	6	404.26	PQQ dependent dehydrogenase, methanol/ethanol family	Hyphomicrobium denitrificans (strain ATCC 51888 / DSM 1869 / NCIB 11706 / TK 0415)
	7	362.71	Flagellin	Leptothrix cholodnii (strain ATCC 51168 / LMG 8142 / SP 6)
	8	354.13	hypothetical protein AY	Gaiella sp. SCGC AG 212 M14
	9	340.41	Methanol dehydrogenase large subunit like protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	10	336.89	PQQ dependent dehydrogenase, methanol/ethanol family	Hyphomicrobium denitrificans (strain ATCC 51888 / DSM 1869 / NCIB 11706 / TK 0415)
	11	334.67	Methanol dehydrogenase large subunit like protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	12	328.65	Ethanolamine utilization protein EutM	Escherichia coli (strain K12)
	13	323.2	Methanol dehydrogenase large subunit like protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	14	309.62	ATP synthase subunit beta	Dyadobacter fermentans (strain ATCC 700827 / DSM 18053 / NS114)
	15	280.84	Formaldehyde activating enzyme	Hyphomicrobium denitrificans (strain ATCC 51888 / DSM 1869 / NCIB 11706 / TK 0415)
	16	252.44	Formaldehyde activating enzyme	Hyphomicrobium denitrificans (strain ATCC 51888 / DSM 1869 / NCIB 11706 / TK 0415)
	17	203.95	ATP synthase subunit beta	Dyadobacter fermentans (strain ATCC 700827 / DSM 18053 / NS114)
	18	194.65	ATP synthase subunit beta	Dyadobacter fermentans (strain ATCC 700827 / DSM 18053 / NS114)
	19	194.2	Methanol dehydrogenase large subunit like protein	Bradyrhizobium diazoefficiens (strain JCM 10833 / IAM 13628 / NBRC 14792 / USDA 110)
	20	188.69	DNA directed RNA polymerase subunit beta	Alkalilimnicola ehrlichii (strain ATCC BAA 1101 / DSM 17681 / MLHE 1)